JOURNAL OF APPLIED PSYCHOLOGY

EDWIN A. FLEISHMAN, EDITOR
American Institutes for Research

JOHN P. CAMPBELL, ASSOCIATE EDITOR
University of Minnesota

CONSULTING EDITORS

EARL A. ALLUISI
Old Dominion University
BERNARD M. BASS
University of Rochester
DONALD E. BROADBENT
Applied Psychology Unit, Medical Research Council, Cambridge, England
PIETER J. D. DRENTH
Free University, Amsterdam, The Netherlands
MARVIN D. DUNNETTE
University of Minnesota
FRANK FRIEDLANDER
Case Western Reserve University
ALBERT S. GLICKMAN
American Institutes for Research, Washington, D. C.
LEONARD V. GORDON
State University of New York at Albany
L. RICHARD HOFFMAN
University of Chicago
PAUL HORST
University of Washington
DAVID KIPNIS
Temple University
EDWIN A. LOCKE
University of Maryland
JOSEPH E. MCGRATH
University of Illinois, Urbana
QUINN MCNEMAR
Stanford University
ROBERT B. MILLER
IBM, Poughkeepsie, New York
WILLIAM G. MOLLENKOPF
Procter & Gamble Company, Cincinnati, Ohio
WILLIAM A. OWENS
University of Georgia
ROBERT PERLOFF
University of Pittsburgh
FRANCIS J. PILGRIM
Chicago, Illinois
LYMAN W. PORTER
University of California, Irvine
ANNE ROE
University of Arizona
STANLEY E. SEASHORE
University of Michigan
MARY L. TENOPYR
AT & T, New York
HARRY TRIANDIS
University of Illinois, Urbana
DON TRUMBO
Pennsylvania State University
J. E. UHLANER
U. S. Army Research Institute for the Behavioral and Social Sciences, Arlington, Virginia

VOLUME 59, 1974

HAROLD P. VAN COTT
Managing Editor

ANITA DE VIVO
Executive Editor and General Manager

DEBORAH S. FELDER
Supervising Technical Editor

SHELLEY M. HAMMOND
Technical Editor

LYNN D. SPIER
Acting Manager, Publications Business Services

ANNE REDMAN
Subscription Manager

$25.00

PUBLISHED BIMONTHLY BY
THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
1200 Seventeenth St., N. W., Washington, D. C. 20036

Copyright © 1974 by the American Psychological Association, Inc.

[illegible] & Psyl. Research
(N. C. E. R. T.)
6·3·81
J 702

ACKNOWLEDGMENTS

I am pleased to list those individuals who, in addition to members of the Editorial Board of Consulting Editors, assisted in the review of manuscripts submitted to the *Journal* between January 1973 and December 1973. I would now like to acknowledge, with appreciation, the contributions of the following individuals.

Edwin A. Fleishman
Editor

Norman M. Abrahams
J. Stacy Adams
Clayton P. Alderfer
Edward F. Alf, Jr.
Lynn R. Anderson
Norman H. Anderson
I. Robert Andrews
William H. Angoff
Philip Ash
William J. Baker
Gerald V. Barrett
Richard S. Barrett
C. J. Bartlett
Albert E. Bartz
Alan R. Bass
Howard J. Baumgartel
Raymond A. Bauer
James A. Dayton
Sandra L. Bem
Leonard Blank
Milton R. Blood
Virginia R. Boehm
Leo Bogart
Edgar F. Borgatta
Thomas J. Bouchard, Jr.
Arthur H. Brayfield
Timothy C. Brock
Gary B. Brumback
Paul C. Buchanan
C. Victor Bunderson
David P. Campbell
Robert E. Carlson
William C. Carrigan
Dorwin P. Cartwright
Ronald P. Carver
Raymond E. Christal
Russell D. Clark III
Norman Cliff
Nancy S. Cole
A. J. Conger
Royer F. Cook
James E. Cowden
Lee J. Cronbach
Ervin W. Curtis
H. Peter Dachler
Edward L. Deci
John M. Digman
Dorothy S. Edwards
Hillel J. Einhorn
Henry C. Ellis
George W. England
James L. Farr
Fred E. Fiedler
R. H. Finn
Martin E. Fishbein
Eli Flyer
Uriel G. Foa
Norman O. Frederiksen
Robert L. Frey
Alinda Friedman
Monroe P. Friedman
Benjamin Fruchter
Wendell R. Garner
Basil S. Georgopoulos
Edwin E. Ghiselli
David C. Glass
Goldine C. Gleser
Milton R. Goldsamt
Bram C. Goldwater
Paul S. Goodman
Barry E. Goodstadt
Michael E. Gordon
George B. Graen
Donald L. Grant
Walter F. Grether
Frederick E. Guedry, Jr.
Leopold W. Gruenfeld
Lester P. Guest
Joan S. Guilford
Robert M. Guion
Edward J. Gunderson
Clifford P. Hahn
Milton D. Hakel
Douglas T. Hall
O. J. Harvey
William H. Helme
John K. Hemphill
Jeanne B. Herman
Paul J. Hoffman
John L. Holland
George P. Huber
Charles L. Hulin
Lloyd G. Humphreys
Bernard P. Indik
Jacob Jacoby
Raymond L. Johnson
James W. Julian
Robert L. Kahn
Raymond A. Katzell

Emanuel Kay
Harold H. Kelley
Wayne K. Kirchner
George R. Klare
Abraham K. Korman
Herbert E. Krugman
Frank J. Landy
Edward E. Lawler
Herbert M. Lefcourt
Jerome P. LeJeune
Russell L. Leonard, Jr.
Jerrold M. Levine
Helen Z. Lopata
Aaron Lowin
George W. McConkie
Ernest J. McCormick
Wilbert J. McKeachie
Louis L. McQuitty
Norman R. F. Maier
Richard D. Mann
Michael S. Matell
Eugene C. Mayfield
Robert K. Merton
Herbert H. Meyer
George Milkovich
James W. Miller
Monroe J. Miller
Angelo Mirabella
Terence R. Mitchell
Charles T. Myers
Allen Newell
David B. Orr
Dean Peabody
Lawrence A. Pervin
Lawrence D. Phillips
E. C. Poulton
Erich P. Prien
Robert D. Pritchard
Roland W. Radloff
James M. Richards, Jr.
Joseph W. Rigney
John R. Rizzo
Karlene H. Roberts
Leonard G. Rorer
Andrew M. Rose
Benson Rosen
Ernst Z. Rothkopf
Julian B. Rotter
Patricia M. Rowe
Edward A. Rundquist
Joseph E. Russo
Virginia E. Schein
Frank L. Schmidt
William A. Scott
Melvin Seeman
Robert Seibel
Saul B. Sells
Jagdish N. Sheth
Harris H. Shettel, Jr.
Robert N. Singer
Patricia Cain Smith
Janet T. Spence
Matilde K. Stephenson
Joseph C. Stevens
Thomas G. Sticht
Ralph M. Stogdill
Arnold S. Tannenbaum
Warren H. Teichner
Robert L. Thorndike
David V. Tiedeman
Enzo R. Valenzi
Neil J. Vidmar
Victor H. Vroom
Michael A. Wallach
John P. Wanous
Edward C. Webster
Karl E. Weick, Jr.
Bernard Weiner
Bernard Weiss
David J. Weiss
Joseph Weitz
William F. Weitzel, Jr.
Kenneth N. Wexley
George R. Wheaton
Gerald H. Whitlock
Allan W. Wicker
Robert T. Wilkinson
Gary A. Yukl
Sheldon Zedeck

CONTENTS OF VOLUME 59

Aaker, David A., and Day, George S. A dynamic model of relationships among advertising, consumer awareness, attitudes, and behavior 281
Allen, John A., Jr., Schroeder, Stephen R., and Ball, Patricia G. Effect of head restriction on drivers' eye movements and errors in simulated dangerous situations 643
Altavela, Julie. *See* Gitter, A. George.
Amos, Moira. *See* Gordon, Ian E.
Andrews, I. R. *See* McClurg, J. M.
Arvey, Richard D. *See* Gordon, Michael E.
Arvey, Richard D., and Hoyle, Joseph C. A Guttman approach to the development of behaviorally based rating scales for systems analysts and programmer/analysts 61
Arvey, Richard D., and Mussio, Stephen J. Job expectations and valences of job rewards for culturally disadvantaged and advantaged clerical employees 230
Ashton, Robert H. Cue utilization and expert judgments: A comparison of independent auditors with other judges 437
Badin, Irwin J. Some moderator influences on relationships between consideration, initiating structure, and organizational criteria 380
Badin, Irwin J. *See* Greenhaus, Jeffrey H.
Baird, Lloyd. *See* Hamner, W. Clay.
Ball, Patricia G. *See* Allen, John A., Jr.
Barak, Azy. *See* Meir, Elchanan I.
Barker, Anne S. *See* Fowler, Robert L.
Barsaloux, Jean. *See* Bouchard, Thomas J., Jr.
Bass, Bernard M., Cascio, Wayne F., and O'Connor, Edward J. Magnitude estimations of expressions of frequency and amount 313
Beatty, Richard W. Supervisory behavior related to job success of hard-core unemployed over a two-year period 38
Bettman, James R. Relationship of information-processing attitude structures to private brand purchasing behavior 79
Biersner, Robert J., and Ryman, David H. Prediction of scuba training performance 519
Bigoness, William J. *See* Hamner, W. Clay.
Blair, Michael O. *See* Wainer, Howard.
Blaiwes, Arthur S. Formats for presenting procedural instructions 683
Blood, Milton R. Spin-offs from behavioral expectation scale procedures 513
Blumenfeld, Warren S., and Sartain, Patricia L. Predicting alumni financial donation 522
Boltwood, Charles E. *See* Cooper, Michael R.
Booth, Jerome M. *See* Lee, Raymond.
Booth, Richard F., and Hoiberg, Anne. Structure and measurement of Marine recruit attitudes 236
Borman, Walter C., and Vallon, W. Robert. A view of what can happen when behavioral expectation scales are developed in one setting and used in another 197
Bouchard, Thomas J., Jr., Barsaloux, Jean, and Drauden, Gail. Brainstorming procedure, group size, and sex as determinants of the problem-solving effectiveness of groups and individuals 135
Bouchard, Thomas J., Jr., Drauden, Gail, and Barsaloux, Jean. A comparison of individual, subgroup, and total group methods of problem solving 226
Boulian, Paul V. *See* Porter, Lyman W.
Braucht, G. N. *See* Searls, Donald J.
Breed, George. *See* Colaiuta, Victoria.
Brodie, B. J. *See* Crocker, L. M.
Burnaska, Robert F., and Hollmann, Thomas D. An empirical comparison of the relative effects of rater response biases on three rating scale formats 307
Butcher, James N., and Ryan, Michael. Personality stability and adjustment to an extreme environment 107
Butler, Richard P. *See* McLaughlin, Gerald W.
Butler, Richard P., and Jaffee, Cabot L. Effects of incentive, feedback, and manner of presenting the feedback on leader behavior 332
Butterfield, D. Anthony, and Farris, George F. The Likert Organizational Profile: Methodological analysis and test of System 4 Theory in Brazil 15
Cafferty, Thomas P., and Streufert, Siegfried. Conflict and attitudes toward the opponent: An application of the Collins and Hoyt attitude change theory to groups in inter-organizational conflict 48
Cascio, Wayne F. *See* Bass, Bernard M.
Cashman, James. *See* Dansereau, Fred.
Chapanis, Alphonse. *See* Weeks, Gerald D.

Clanton, Earl S., III. *See* Johnson, W. Russell.
Coatney, Richard P. *See* Rothkopf, Ernst Z.
Cohen, Jeffrey, and Lefkowitz, Joel. Development of a biographical inventory blank to predict faking on personality tests 404
Colaiuta, Victoria, and Breed, George. Development of scales to measure attitudes toward marijuana and marijuana users 398
Cooper, Michael R., Boltwood, Charles E. and Wherry, Robert J., Sr. A factor analysis of air passenger reactions to skyjacking and airport security measures as related to personal characteristics and alternatives to flying 365
Cooper, Michael R., and Wood, Michael T. Effects of member participation and commitment in group decision making on influence, satisfaction, and decision riskiness 127
Crocker, L. M., and Brodie, B. J. Development of a scale to assess student nurses' views of the professional nursing role 233
Csoka, Louis S. A relationship between leader intelligence and leader rated effectiveness 43
Daffron, William C., Jr. *See* Gordon, Michael E.
Dansereau, Fred, Cashman, James, and Graen, George. Expectancy as a moderator of the relationship between job attitudes and turnover 228
Day, George S. *See* Aaker, David A.
Dermer, Jerry. *See* Evans, Martin G.
Dillman, Don A., and Frey, James H. Contribution of personalization to mail questionnaire response as an element of a previously tested method 297
Donnelly, James H., Jr., and Etzel, Michael J. Attempting to operationalize product newness: A reply to Ostlund and Tellefsen 761
Dowd, Patrick J. Sleep deprivation effects on the vestibular habituation process 748
Drauden, Gail. *See* Bouchard, Thomas J., Jr.
Dreger, Ralph Mason. *See* Kennedy, Pattye Weaver.
Dunnette, Marvin D. *See* Taylor, Ronald N.
Eden, Dov, and Leviatan, Uri. Farm and factory in the kibbutz: A study in agrico-industrial psychology 596
Edwards, Shirley. *See* McCarrey, Michael W.
Einhorn, Hillel J. Expert judgment: Some necessary conditions and an example 562
Ellison, Robert L. *See* James, Lawrence R.
England, George W., and Lee, Raymond. The relationship between managerial values and managerial success in the United States, Japan, India, and Australia 411
Esposito, James P., and Richards, Herbert C. Dogmatism and the congruence between self-reported job preference and performance among school supervisors 389
Etzel, Michael J. *See* Donnelly, James H., Jr.
Etzel, Michael J., and Walker, Bruce J. Effects of alternative follow-up procedures on mail survey response rates 219
Evans, Martin G. Extensions of a path-goal theory of motivation 172
Evans, Martin G., and Dermer, Jerry. What does the Least Preferred Co-worker Scale really measure? A cognitive interpretation 202
Faggen-Steckler, Jane. *See* Gross, Alan L.
Farber, Martin G. *See* Kelly, Rita Mae.
Farris, George F. *See* Butterfield, D. Anthony.
Feldman, Jack. Race, economic class, and the intention to work: Some normative and attitudinal correlates 179
Feldman, Jack M. Note on the utility of certainty weights in expectancy theory 727
Feldman, Jack M. *See* Triandis, Harry C.
Fessler, Paul K. *See* Laughery, Kenneth R.
Fhanér, Gunilla, and Hane, Monica. Seat belts: Relations between beliefs, attitude, and use . 472
Fleishman, Edwin A. *See* Theologus, George C.
Fodor, Eugene M. Disparagement by a subordinate as an influence on the use of power .. 652
Fossum, John. *See* Moore, Michael L.
Fowler, Robert L., and Barker, Anne S. Effectiveness of highlighting for retention of text material 358
Fox, David G. *See* James, Lawrence R.
Frey, James H. *See* Dillman, Don A.
Frieling, E., Kannheiser, W., and Lindberg, R. Some results with the German form of the Position Analysis Questionnaire (PAQ) 741
Fuller, Carol. Effect of anonymity on return rate and response bias in a mail survey 292
Gamboa, Victor U. *See* Pedalino, Ed.
Gannon, Martin J., and Paine, Frank T. Unity of command and job attitudes of managers in a bureaucratic organization 392
Geacintov, Tamara, and Peavler, W. Scott. Pupillography in industrial fatigue assessment .. 213

Gitter, A. George, Altavela, Julie, and Mostofsky, David I. Effect of sex, religion, and ethnicity on occupational status perception 96

Gleason, Edwin M. *See* Huck, Schuyler W.

Gordon, Francine E., and Hall, Douglas T. Self-image and stereotypes of femininity: Their relationship to women's role conflicts and coping 241

Gordon, Ian E., and Amos, Moira. Checking groups of letters 354

Gordon, Michael E., Arvey, Richard D., Daffron, William C., Jr., and Umberger, Dennis L. Racial differences in the impact of mathematics training at a manpower development program 253

Graen, George. *See* Dansereau, Fred.

Greene, Charles N. *See* Organ, Dennis W.

Greenhaus, Jeffrey H., and Badin, Irwin J. Self-esteem, performance, and satisfaction: Some tests of a theory 722

Gross, Alan L., Faggen-Steckler, Jane, and McCarthy, Karen. Statistical procedures for evaluating the practical utility of a moderator approach to prediction 578

Gruenfeld, L. W., and Weissenberg, P. Relationship between supervisory cognitive style and social orientation 386

Gruvaeus, Gunnar T. *See* Wainer, Howard.

Gunderson, E. K. Eric. *See* McDonald, Blair W.

Gustanski, Jerry. *See* Sanders, Mark.

Hadley, Robert G. Comment on Locke's satisfiers and dissatisfiers among white-collar and blue-collar employees 507

Hagen, Eileen L. *See* Kesselman, Gerald A.

Hakel, Milton D. *See* London, Manuel.

Hall, Douglas T. *See* Gordon, Francine E.

Hamner, W. Clay, Kim, Jay S., Baird, Lloyd and Bigoness, William J. Race and sex as determinants of ratings by potential employers in a simulated work-sampling task .. 705

Hamner, W. Clay, and Tosi, Henry L. Relationship of role conflict and role ambiguity to job involvement measures 497

Hane, Monica. *See* Fhanér, Gunilla.

Harlow, Dorothy N. *See* Miner, John B.

Harris, Thomas C., and Locke, Edwin A. Replication of white-collar–blue-collar differences in sources of satisfaction and dissatisfaction 369

Harvey, William M. *See* Triandis, Harry C.

Hegarty, W. Harvey. Using subordinate ratings to elicit behavioral changes in supervisors .. 764

Heisler, W. J. A performance correlate of personal control beliefs in an organizational context 504

Helms, Samuel T. *See* Nafziger, Dean H.

Hendon, Donald W. Mechano-content analysis as an aid to predicting effects of advertising microcommunications 773

Heneman, Herbert G., III. Comparisons of self- and superior ratings of managerial performance 638

Herold, David M. Interaction of subordinate and leader characteristics in moderating the consideration-satisfaction relationship 649

Hill, James W. *See* Miner, John B.

Hillery, Joseph M., and Wexley, Kenneth N. Participation effects in appraisal interviews conducted in a training situation 168

Hoiberg, Anne. *See* Booth, Richard F.

Hollmann, Thomas D. *See* Burnaska, Robert F.

Hoover, Thomas O. *See* Wagner, Edwin E.

Hoyle, Joseph C. *See* Arvey, Richard D.

Hoyt, G. C. *See* Schoner, Bertram.

Hsu, Chris Ching-Yang, and Newton, Richard R. Relation between foremen's leadership attitudes and the skill level of their work groups 771

Huck, Schuyler W., and Gleason, Edwin M. Using monetary inducements to increase response rates from mailed surveys: A replication and extension of previous research 222

Ilgen, Daniel R., and Seely, William. Realistic expectations as an aid in reducing voluntary resignations 452

Imparato, Nicholas. *See* Zedeck, Sheldon.

Ivancevich, John M. Effects of the shorter workweek on selected satisfaction and performance measures 717

Jacoby, Jacob. *See* Kaplan, Leon B.

Jacoby, Jacob. *See* Szybillo, George J.

Jaffee, Cabot L. *See* Butler, Richard P.

James, Lawrence R., Ellison, Robert L., Fox, David G., and Taylor, Calvin W. Prediction of artistic performance from biographical data 84

Jayson, Jill K. *See* Maas, James B.

Jennings, J. Richard, Rose, R. M., and Kreuz, Leo. Stress and performance during and after Officer Candidate School 500

Jerdee, Thomas H. *See* Rosen, Benson.

Jerdee, Thomas H., and Rosen, Benson. Effects of opportunity to communicate and visibility of individual decisions on behavior in the common interest 712

Johnson, James C. *See* Kaufman, Gary G.

Johnson, W. Russell, Sieveking, Nicholas A., and Clanton, Earl S., III. Effects of alternative positioning of open-ended questions in multiple-choice questionnaires 776

Johnston, H. Russell, Jr. Some personality correlates of the relationships between individuals and organizations 623

Kannheiser, W. *See* Frieling, E.

Kaplan, Leon B., Szybillo, George J., and Jacoby, Jacob. Components of perceived risk in product purchase: A cross-validation 287

Kaufman, Gary G., and Johnson, James C. Scaling peer ratings: An examination of the differential validities of positive and negative nominations 302

Kaufman, H. G. Relationship of early work challenge to job performance, professional contributions, and competence of engineers 377

Kauma, Claudia E. *See* Wicker, Allan W.

Keiper, Charles G. *See* Regina, Edmund G.

Kelly, Michael J. *See* Weeks, Gerald D.

Kelly, Rita Mae, and Farber, Martin G. Identifying responsive inner-city policemen 259

Kennedy, Pattye Weaver, and Dreger, Ralph Mason. Development of criterion measures of overseas missionary performance 69

Kesselman, Gerald A., Wood, Michael T., and Hagen, Eileen L. Relationships between performance and satisfaction under contingent and noncontingent reward systems 374

Kilmann, Ralph H. Participative management in the college classroom 337

Kim, Jay S. *See* Hamner, W. Clay.

Kinne, Sydney B., III. *See* Latham, Gary P.

Kleiber, Douglas A. *See* Maas, James B.

Klemmer, E. T., and Stocker, L. P. Effects of grouping of printed digits on forced-paced manual entry performance 675

Klimoski, Richard J., and London, Manuel. Role of the rater in performance appraisal 445

Korman, Abraham K. A disguised measure of civil rights attitudes 239

Krausz, Moshe. *See* Zedeck, Sheldon.

Kreuz, Leo. *See* Jennings, J. Richard.

Kugelmass, Sol. *See* Lieblich, Israel.

Latham, Gary P., and Kinne, Sydney B., III. Improving job performance through training in goal setting 187

Laughery, Kenneth R., Fessler, Paul K., Lenorovitz, David R., and Yoblick, Darryl A. Time delay and similarity effects in facial recognition 490

Lawton, Mike. *See* Sanders, Mark.

Lee, Alec J. *See* Moore, Larry F.

Lee, Raymond. *See* England, George W.

Lee, Raymond, and Booth, Jerome M. A utility analysis of a weighted application blank designed to predict turnover for clerical employees 516

Lefkowitz, Joel. *See* Cohen, Jeffrey.

Lenorovitz, David R. *See* Laughery, Kenneth R.

Leviatan, Uri. *See* Eden, Dov.

Lichtenstein, Jacob H., and Saucer, Rayford T. Visual dependency in the erect and supine positions 529

Lieblich, Israel, Naftali, Gideon, Shmueli, Joseph, and Kugelmass, Sol. Efficiency of GSR detection of information with repeated presentation of series of stimuli in two motivational states 113

Lindberg, R. *See* Frieling, E.

Lindvall, Thomas, and Svensson, Leif T. Equal unpleasantness matching of malodorous substances in the community 264

Locke, Edwin A. *See* Harris, Thomas C.

Locke, Edwin A., and Whiting, Roman J. Sources of satisfaction and dissatisfaction among solid waste management employees 145

London, Manuel. *See* Klimoski, Richard J.

London, Manuel, and Hakel, Milton D. Effects of applicant stereotypes, order, and information on interview impressions 157

Louviere, Jordan J. Predicting the evaluation of real stimulus objects from abstract evaluation of their attributes: The case of trout streams 572

Louviere, Jordan J. *See* Norman, Kent L.

Maas, James B., Jayson, Jill K., and Kleiber, Douglas A. Effects of spectral differences in illumination on fatigue 524

Matteson, Michael T. Type of transmittal letter and questionnaire color as two variables influencing response rates in a mail survey 535

McCarrey, Michael W., Peterson, Lyle, Edwards, Shirley, and Von Kulmiz, Paul. Landscape office attitudes: Reflections of perceived degree of control over transactions with the environment 401

McCarthy, Karen. *See* Gross, Alan L.

McClurg, J. M., and Andrews, I. R. A consumer profile analysis of the self-service gasoline customer 119

McCullough, J. Lee, and Ostrom, Thomas M. Repetition of highly similar messages and attitude change 395

McDonald, Blair W., and Gunderson, E. K. Eric. Correlates of job satisfaction in naval environments 371

McKelvey, Robert E. *See* Regina, Edmund G.

McLaughlin, Gerald W., and Butler, Richard P. Use of weighting and data on alternative careers to predict career retention and commitment 287

Mecham, Robert C. *See* Robinson, David D.

Meir, Elchanan I., and Barak, Azy. Pervasiveness of the relationship between intrinsic–extrinsic needs and persistence at work 103

Miller, Edwin. *See* Moore, Michael L.

Miner, John B., Rizzo, John R., Harlow, Dorothy N., and Hill, James W. Role motivation theory of managerial effectiveness in simulated organizations of varying degrees of structure 31

Miskel, Cecil. Intrinsic, extrinsic, and risk propensity factors in the work attitudes of teachers, educational administrators, and business managers 339

Miskimins, R. W. *See* Searls, Donald J.

Moore, Larry F., and Lee, Alec J. Comparability of interviewer, group, and individual interview ratings 163

Moore, Michael L., Miller, Edwin, and Fossum, John. Predictors of managerial career expectations 90

Morrison, Robert F., and Sebald, Maria-Luise. Personal characteristics differentiating female executive from female nonexecutive personnel 656

Mostofsky, David I. *See* Gitter, A. George.

Mowday, Richard T. *See* Porter, Lyman W.

Mussio, Stephen J. *See* Arvey, Richard D.

Naftali, Gideon. *See* Lieblich, Israel.

Nafziger, Dean H., and Helms, Samuel T. Cluster analyses of interest inventory scales as tests of Holland's occupational classification 344

Napoliello, Marie. *See* Oswalt, Robert M.

Newman, John E. Predicting absenteeism and turnover: A field comparison of Fishbein's model and traditional job attitude measures 610

Newton, Richard R. *See* Hsu, Chris Ching-Yang.

Norman, Kent L., and Louviere, Jordan J. Integration of attributes in bus transportation: Two modeling approaches 753

O'Connor, Edward J. *See* Bass, Bernard M.

Oleno, Timothy. *See* Zedeck, Sheldon.

Olson, Paul L. Aspects of driving performance as a function of field dependence 192

O'Reilly, Charles A., III. *See* Roberts, Karlene H.

Organ, Dennis W., and Greene, Charles N. Role ambiguity, locus of control, and work satisfaction 101

Ostlund, Lyman E., and Tellefsen, Brynulf. Relationship between consumers' category width and trial of new products: A reappraisal 759

Ostrom, Thomas M. *See* McCullough, J. Lee

Oswalt, Robert M., and Napoliello, Marie. Motivations of blood donors and nondonors ... 122

Paine, Frank T. *See* Gannon, Martin J.

Peavler, W. Scott. *See* Geacintov, Tamara.

Pedalino, Ed, and Gamboa, Victor U. Behavior modification and absenteeism: Intervention in one industrial setting 694

Peterson, Lyle. *See* McCarrey, Michael W.

Petty, M. M. Relative effectiveness of four combinations of oral and written presentations of job related information to disadvantaged trainees 105

Petty, M. M., and Pryor, Norman M. A note on the predictive validity of initiating structure and consideration in ROTC training 383

Pfeifer, C. Michael, Jr., and Schneider, Benjamin. University climate perceptions by black and white students 660

Porter, Lyman W., Steers, Richard M., Mowday, Richard T., and Boulian, Paul V. Organizational commitment, job satisfaction, and turnover among psychiatric technicians 603

Pryor, Norman M. *See* Petty, M. M.

Regina, Edmund G., Smith, Gene M., Keiper, Charles G., and McKelvey, Robert E. Effects of caffeine on alertness in simulated automobile driving 483

Richards, Herbert C. *See* Esposito, James P.

Richards, James M., Jr. Some psychosocial consequences of a change to a replacement birthrate in the United States 1

Rizzo, John R. *See* Miner, John B.

Roberts, Karlene H., and O'Reilly, Charles A., III. Measuring organizational communication 321

Robinson, David D., Wahlstrom, Owen W., and Mecham, Robert C. Comparison of job evaluation methods: A "policy-capturing" approach using the Position Analysis Questionnaire 633

Rollo, James. *See* Smith, Patricia Cain.

Rose, Gerald L. *See* Schoner, Bertram.

Rose, R. M. *See* Jennings, J. Richard.

Rosen, Benson. *See* Jerdee, Thomas H.

Rosen, Benson, and Jerdee, Thomas H. Effects of applicant's sex and difficulty of job on evaluations of candidates for managerial positions 511

Rosen, Benson, and Jerdee, Thomas H. Factors influencing disciplinary judgments 327

Rosen, Benson, and Jerdee, Thomas H. Influence of sex role stereotypes on personnel decisions 9

Rothkopf, Ernst Z., and Coatney, Richard P. Effects of readability of context passages on subsequent inspection rates 679

Ryan, Michael. *See* Butcher, James N.

Ryman, David H. *See* Biersner, Robert J.

Sanders, Mark, Gustanski, Jerry, and Lawton, Mike. Effect of ambient illumination on noise level of groups 527

Sartain, Patricia L. *See* Blumenfeld, Warren S.

Sashkin, Marshall, Taylor, F. Carter, and Tripathi, Rama C. An analysis of situational moderating effects on the relationships between Least Preferred Co-worker and other psychological measures 731

Saucer, Rayford T. *See* Lichtenstein, Jacob H.

Scherf, Gerhard W. H. Consumer dissatisfaction as a function of dissatisfaction with interpersonal relationships 465

Schmidt, Frank L. Probability and utility assumptions underlying use of the Strong Vocational Interest Blank 456

Schneider, Benjamin. *See* Pfeifer, C. Michael, Jr.

Schneiderman, Mark L. *See* Wiener, Yoash.

Schoenfeldt, Lyle F. Utilization of manpower: Development and evaluation of an assessment-classification model for matching individuals with jobs 583

Schoner, Bertram, Hoyt, G. C., and Rose, Gerald L. Quality of decisions: Individuals versus real and synthetic groups 424

Schroeder, Stephen R. *See* Allen, John A., Jr.

Searls, Donald J., Braucht, G. N., and Miskimins, R. W. Work values of the chronically unemployed 93

Sebald, Maria-Luise. *See* Morrison, Robert F.

Seely, William. *See* Ilgen, Daniel R.

Shmueli, Joseph. *See* Lieblich, Israel.

Sieveking, Nicholas A. *See* Johnson, W. Russell.

Sims, Henry P., Jr. *See* Szilagyi, Andrew D.

Smith, Gene M. *See* Regina, Edmund G.

Smith, Olin W. *See* Smith, Patricia Cain.

Smith, Patricia Cain, Smith, Olin W., and Rollo, James. Factor structure for blacks and whites of the Job Descriptive Index and its discrimination of job satisfaction 99

South, John C. Achievement motivation among managers of small businesses, corporation managers, and business students 509

Spadaro, Robert N. Some folkways of political leaders: The rules of the game 125

Sperry, Len. Effects of expectation, social class, and experience on in-service teacher behavior in small groups 244

Srinivasan, V. *See* Weinstein, Alan G.

Steers, Richard M. *See* Porter, Lyman W.

Stocker, L. P. *See* Klemmer, E. T.

Street, Warren R. Brainstorming by individuals, coacting and interacting groups 433

Streufert, Siegfried. *See* Cafferty, Thomas P.

Svensson, Leif T. *See* Lindvall, Thomas.

Swezey, Robert W. Brightness contrast effects on recall of projected highway sign-type stimulus material 408
Szilagyi, Andrew D., and Sims, Henry P., Jr. Cross-sample stability of the Supervisory Behavior Description questionnaire 767
Szybillo, George J. *See* Kaplan, Leon B.
Szybillo, George J., and Jacoby, Jacob. Effects of different levels of integration on advertising preference and intention to purchase 274
Szybillo, George J., and Jacoby, Jacob. Intrinsic versus extrinsic cues as determinants of perceived product quality 74
Taylor, Calvin W. *See* James, Lawrence R.
Taylor, F. Carter. *See* Sashkin, Marshall.
Taylor, Ronald N., and Dunnette, Marvin D. Influence of dogmatism, risk-taking propensity, and intelligence on decision-making strategies for a sample of industrial managers 420
Tellefsen, Brynulf. *See* Ostlund, Lyman E.
Theologus, George C., Wheaton, George R., and Fleishman, Edwin A. Effects of intermittent, moderate intensity noise stress on human performance 539
Tosi, Henry L. *See* Hamner, W. Clay.
Triandis, Harry C., Feldman, Jack M., Weldon, David E., and Harvey, William M. Designing preemployment training for the hard to employ: A cross-cultural psychological approach 687
Tripathi, Rama C. *See* Sashkin, Marshall.
Umberger, Dennis L. *See* Gordon, Michael E.
Vallon, W. Robert. *See* Borman, Walter C.
Veiga, John F. Getting the mail questionnaire returned: Some practical research considerations 217
Voevodsky, John. Evaluation of a deceleration warning light for reducing rear-end automobile collisions 270
Von Kulmiz, Paul. *See* McCarrey, Michael W.
Wagner, Edwin E., and Hoover, Thomas O. Effect of terminal ranking error on meaningful occupational choice 247
Wagner, Edwin E., and Hoover, Thomas O. The influence of technical knowledge on position error in ranking 406
Wahlstrom, Owen W. *See* Robinson, David D.
Wainer, Howard, Gruvaeus, Gunnar T., Blair, Michael O., and Zill, Nicholas, II. Using variance as a discriminator in lie detection 110
Walker, Bruce J. *See* Etzel, Michael J.
Wanous, John P. A causal-correlational analysis of the job satisfaction and performance relationship 139
Wanous, John P. Individual differences and reactions to job characteristics 616
Weeks, Gerald D., Kelly, Michael J., and Chapanis, Alphonse. Studies in interactive communciation: V. Cooperative problem solving by skilled and unskilled typists in a teletypewriter mode 665
Weinstein, Alan G., and Srinivasan, V. Predicting managerial success of Master of Business Administration (MBA) graduates 207
Weinstein, Neil D. Effect of noise on intellectual performance 548
Weissenberg, P. *See* Gruenfeld, L. W.
Weldon, David E. *See* Triandis, Harry C.
Wexley, Kenneth N. *See* Hillery, Joseph M.
Wheaton, George R. *See* Theologus, George C.
Wherry, Robert J., Sr. *See* Cooper, Michael R.
Whiting, Roman J. *See* Locke, Edwin A.
Wicker, Allan W., and Kauma, Claudia E. Effects of a merger of a small and a large organization on members' behaviors and experiences 24
Wiener, Yoash, and Schneiderman, Mark L. Use of job information as a criterion in employment decisions of interviewers 699
Witt, William. Effects of numerical data upon reading time and recall of scientific writing ... 532
Wood, Donald A. Effect of worker orientation differences on job attitude correlates 54
Wood, Michael T. *See* Cooper, Michael R.
Wood, Michael T. *See* Kesselman, Gerald A.
Woodside, Arch G. Relation of price to perception of quality of new products 116
Wright, Peter. The harassed decision maker: Time pressures, distractions, and the use of evidence 555
Yoblick, Darryl A. *See* Laughery, Kenneth R.
Zedeck, Sheldon, Imparato, Nicholas, Krausz, Moshe, and Oleno, Timothy. Development of behaviorally anchored rating scales as a function of organizational level 249
Zill, Nicholas, II. *See* Wainer, Howard.

JOURNAL OF APPLIED PSYCHOLOGY

Copyright © 1974 by the American Psychological Association, Inc.

February 1974 Vol. 59, No. 1

ARTICLES

Some Psychosocial Consequences of a Change to a Replacement Birthrate in the United States *James M. Richards, Jr.* 1

Influence of Sex Role Stereotypes on Personnel Decisions *Benson Rosen and Thomas H. Jerdee* 9

The Likert Organizational Profile: Methodological Analysis and Test of System 4 Theory in Brazil *D. Anthony Butterfield and George F. Farris* 15

Effects of a Merger of a Small and a Large Organization on Members' Behaviors and Experiences *Allan W. Wicker and Claudia E. Kauma* 24

Role Motivation Theory of Managerial Effectiveness in Simulated Organizations of Varying Degrees of Structure *John B. Miner, John R. Rizzo, Dorothy N. Harlow, and James W. Hill* 31

Supervisory Behavior Related to Job Success of Hard-Core Unemployed over a Two-Year Period *Richard W. Beatty* 38

A Relationship between Leader Intelligence and Leader Rated Effectiveness *Louis S. Csoka* 43

Conflict and Attitudes toward the Opponent: An Application of the Collins and Hoyt Attitude Change Theory to Groups in Interorganizational Conflict *Thomas P. Cafferty and Siegfried Streufert* 48

Effect of Worker Orientation Differences on Job Attitude Correlates *Donald A. Wood* 54

A Guttman Approach to the Development of Behaviorally Based Rating Scales for Systems Analysts and Programmer/Analysts *Richard D. Arvey and Joseph C. Hoyle* 61

Development of Criterion Measures of Overseas Missionary Performance *Pattye Weaver Kennedy and Ralph Mason Dreger* 69

Intrinsic Versus Extrinsic Cues as Determinants of Perceived Product Quality *George J. Szybillo and Jacob Jacoby* 74

Relationship of Information-Processing Attitude Structures to Private Brand Purchasing Behavior *James R. Bettman* 79

SHORT NOTES

Prediction of Artistic Performance from Biographical Data *Lawrence R. James, Robert L. Ellison, David G. Fox, and Calvin W. Taylor* 84

Use of Weighting and Data on Alternative Careers to Predict Career Retention and Commitment *Gerald W. McLaughlin and Richard P. Butler* 87

Predictors of Managerial Career Expectations *Michael L. Moore, Edwin Miller, and John Fossum* 90

Work Values of the Chronically Unemployed *Donald J. Searls, G. N. Braucht, and R. W. Miskimins* 93

Diary No. 886(3)
Date 16/9/74

Effect of Sex, Religion, and Ethnicity on Occupational Status Perception
A. George Gitter, Julie Altavela, and David I. Mostofsky 96

Factor Structure for Blacks and Whites of the Job Descriptive Index and Its Discrimination of Job Satisfaction *Patricia Cain Smith, Olin W. Smith, and James Rollo* 99

Role Ambiguity, Locus of Control, and Work Satisfaction
Dennis W. Organ and Charles N. Greene 101

Pervasiveness of the Relationship between Intrinsic–Extrinsic Needs and Persistence at Work *Elchanan I. Meir and Azy Barak* 103

Relative Effectiveness of Four Combinations of Oral and Written Presentations of Job-Related Information to Disadvantaged Trainees *M. M. Petty* 105

Personality Stability and Adjustment to an Extreme Environment
James N. Butcher and Michael Ryan 107

Using Variance As a Discriminator in Lie Detection
Howard Wainer, Gunnar T. Gruvaeus, Michael O. Blair, and Nicholas Zill II 110

Efficiency of GSR Detection of Information with Repeated Presentation of Series of Stimuli in Two Motivational States
Israel Lieblich, Gideon Naftali, Joseph Shmueli, and Sol Kugelmass 113

Relation of Price to Perception of Quality of New Products *Arch G. Woodside* 116

A Consumer Profile Analysis of the Self-Service Gasoline Customer
J. M. McClurg and I. R. Andrews 119

Motivations of Blood Donors and Nondonors *Robert M. Oswalt and Marie Napoliello* 122

Some Folkways of Political Leaders: The Rules of the Game *Robert N. Spadaro* 125

LIST OF MANUSCRIPTS ACCEPTED 130

INSTRUCTIONS TO AUTHORS

Articles must possess: (*a*) conciseness; (*b*) unambiguous and simple vocabulary; (*c*) conformity to accepted technical style in tables, terminology, and references; and (*d*) conclusions that are clearly related to the evidence presented. Authors must progress from a statement of problem or purpose, through analysis of evidence, to conclusions and implications.

Criteria for evaluation include: (*a*) significance of contribution, (*b*) technical adequacy, (*c*) appropriateness for the *Journal of Applied Psychology,* and (*d*) clarity of presentation.

See the *Publication Manual of the American Psychological Association* (1967 Revision), $1.50, for the format of manuscripts. Articles not prepared in this manner cannot be reviewed. Note sections on typing (p. 48, *all copy must be double-spaced*), tables, figures, and references. The first author's last name must be typed on the upper left-hand corner of each page, except for blind reviews. Do not abbreviate, for example, journal titles in reference lists, the word "subject," names of groups, etc. Each manuscript copy must include an abstract of 100–120 words typed on a separate sheet of paper.

Optional blind review must be specifically requested when a manuscript is first submitted. The author's name and affiliation must appear only on a separate title page, included with each copy of the manuscript. Footnotes containing the identity of the author or his affiliation must be on separate pages.

Check recent issues of the *Journal of Applied Psychology* for approximate length of regular articles. (One printed page equals three double-spaced manuscript pages.) A few longer articles of special significance are printed as monographs. Short Notes feature brief reports on studies such as those involving some methodological contribution or important replication.

The original drawing of a figure or an 8 × 10 inch glossy print must be submitted with the original copy of the manuscript. Duplicate copies may be photographic or pencil drawn. Figures *must* be hand-lettered.

Supplementary materials formerly deposited with the National Auxiliary Publications Service need no longer be submitted with the manuscript. Retain supporting and raw data for at least 5 years after publication of the article.

Journal of Applied Psychology
1974, Vol. 59, No. 1, 1-8

SOME PSYCHOSOCIAL CONSEQUENCES OF A CHANGE TO A REPLACEMENT BIRTHRATE IN THE UNITED STATES

JAMES M. RICHARDS, JR.[1]
The Johns Hopkins University

Consequences of a change to a replacement birthrate and a stable population would include changes in the population age distribution. The present study employed estimates of such age distribution changes to estimate changes in several psychological characteristics of the United States population, including average abilities, personality patterns, and incidence of important life transitions. Results indicated that these age distribution changes alone would produce no major changes in the average psychological characteristics of the population of the United States. Therefore, age distribution-related psychological changes do not appear to pose any serious objection to a replacement birthrate in the United States.

This study assumes the essential correctness of the view that human survival and the opportunity for a decent life for all depend on finding solutions to the interrelated problems of stabilizing population, increasing food supply, and living more in harmony with the natural environment and available resources (Ehrlich & Ehrlich, 1972; Hutchinson, 1969; NAS–NRC Committee on Resources and Man, 1969; Kirk, 1973; Looft, 1971; Meadows, Meadows, Randers, & Behrens, 1972). Because most proponents of this view regard continuing population growth as a genuine and critical emergency, and because the most plausible way to stabilize population is by first attaining a replacement birthrate,[2] most of their efforts have been devoted to understanding fertility (Fawcett, 1973; Pohlman, 1969) and to reducing the birthrate. Most recent trends in the United States, however, suggest that a birthrate at (or below) replacement level is likely in the near future. Therefore, it seems appropriate for advocates of such a transition to devote some effort to anticipating the consequences of this change in order to minimize any socially destructive consequences. Knowledge of such consequences should contribute to the scientific understanding of the psychology of population (American Psychological Association Task Force, 1972).

The present study is intended to be one step in such anticipation of consequences. Specifically, a replacement birthrate would produce gradual changes in the population age distribution. The purpose of this research was to estimate the changes in several psychological characteristics of the United States population, including average abilities, personality patterns, and incidence of important life transitions, that would occur as a consequence of such changes in the age distribution.

METHOD

Using the 1969 age distribution as a base, age distributions were computed at five-year intervals for 150 years after attainment of a replacement birthrate.[3] These age distributions were combined with estimates of the relationship between age and various psychological characteristics to yield estimates of population psychological characteristics for the same five-year intervals. For each characteristic, the general procedure was (*a*) to multiply the typical score for people in each of several age ranges by the number of people in that age range in a given year, (*b*) to sum across age ranges, and (*c*) to divide by the total number of people in the population in that year. These computations were severely handicapped by the paucity of hard data about the relationship between age through the entire life span and *any* psychological characteristics (Bloom, 1964).

---

[1] Requests for reprints should be sent to James M. Richards, Jr., Center For Social Organization of Schools, The Johns Hopkins University, Baltimore, Maryland 21218.

[2] A stable population could also be attained, for example, by having the number of births each year exactly equal the number of deaths (Frejka, 1968). This possibility seems quite improbable given the current age distribution of the United States because the birthrate would be low when many women are in the prime child-bearing years and high when few women are in those years.

[3] Tables and figures, however, show results only to the point at which values stabilize.

By deliberate choice, the calculations involved a small number of assumptions or parameter estimates; that is, the calculations involved only estimated age distributions and estimated relationships between age and various psychological characteristics. The effect of such social changes as increases in average educational level were ignored. This procedure assumed that there is an inherent maturational relationship between age and these psychological characteristics. Such an assumption appears essential to the purposes of this study.

In order to compute the age distributions, birthrates specific to women of the United States in various age groups were derived by assuming that these women continue to have their children at the same relative ages, but give birth to only enough female children to yield a net reproduction rate of 1.00. It was also assumed that the death rates for the entire United States population sank to the level of the white population's death rates at the same time a replacement birthrate was attained. (For ages above 70, death rates of the white population were estimated from overall rates.) The use of the white population's death rates for the entire United States population was a partial response to the charge that measures to stabilize population are genocidal. Advocates of stabilization may react emotionally to such maligning of their motives, but it is important to consider what serious arguments may underlie the rhetoric. It is clearly true that death rates for such groups as blacks, Chicanos, and American Indians are substantially higher than death rates for whites. If death rates for these groups stayed the same while birthrates for women of specific ages sank to what would be a replacement rate for whites, the consequences would in fact be gradual disappearance of these groups from the population. In other words, such differential death rates provide clear cut evidence of continuing social injustice in the United States; and it is important to attain both stable population and social justice. The use of death rates of the white population provides a minimal demographic definition of social justice for this study. To facilitate computation these birth and death rates were expressed as probabilities, the age distributions were calculated through simple probability multiplication, and the effects of immigration and emigration were not considered. Therefore the age distributions could differ slightly from those that could be produced by a more complex mathematical model for demography (Coale, 1972).

RESULTS AND DISCUSSION

Demographic changes. The age distributions in various years after attainment of a replacement birthrate are summarized in

TABLE 1

ESTIMATED AGE DISTRIBUTIONS AT FIVE-YEAR INTERVALS ASSUMING A REPLACEMENT BIRTHRATE WAS ATTAINED IN 1969

Year	Age distributions															
	Males								Females							
	0–9	10–19	20–29	30–39	40–49	50–59	60–69	70–up	0–9	10–19	20–29	30–39	40–49	50–59	60–69	70–up
1969	19.8	19.9	15.0	11.1	11.8	10.0	7.0	5.2	19.0	19.2	14.9	11.4	12.5	10.9	8.1	7.5
1974	16.8	21.0	17.6	12.2	11.2	10.6	7.6	5.3	16.1	20.3	17.3	12.4	11.7	11.7	9.0	7.5
1979	16.4	19.7	19.6	14.7	10.8	10.8	8.0	5.6	15.7	18.9	19.1	14.7	11.2	11.9	9.8	7.9
1984	18.3	16.7	20.7	17.3	11.8	10.2	8.5	6.1	17.6	16.1	20.2	17.1	12.1	11.2	10.6	8.7
1989	19.7	16.3	19.4	19.2	14.2	9.8	8.6	6.4	18.8	15.7	18.8	19.0	14.5	10.7	10.8	9.4
1994	19.6	18.3	16.5	20.3	16.7	10.8	8.2	6.9	18.8	17.5	16.0	20.0	16.8	11.6	10.1	10.2
1999	18.5	19.6	16.1	19.0	18.6	13.1	7.9	7.0	17.8	18.8	15.6	18.7	18.6	13.9	9.7	10.5
2004	17.7	19.5	18.0	16.2	19.7	15.3	8.7	6.8	17.0	18.7	17.4	15.9	19.6	16.1	10.5	10.3
2009	17.9	18.4	19.3	15.8	18.4	17.0	10.5	6.7	17.2	17.7	18.7	15.4	18.3	17.8	12.5	10.0
2014	18.6	17.6	19.2	17.7	15.7	18.0	12.3	7.1	17.8	16.9	18.6	17.3	15.6	18.8	14.5	10.5
2019	19.0	17.8	18.2	18.9	15.3	16.8	13.6	8.2	18.2	17.1	17.6	18.5	15.2	17.5	16.1	11.9
2024	18.8	18.5	17.4	18.9	17.1	14.3	14.4	9.5	18.1	17.7	16.8	18.5	17.0	14.9	16.9	13.6
2029	18.4	18.9	17.6	17.9	18.3	14.0	13.4	10.7	17.7	18.1	17.0	17.5	18.2	14.5	15.8	15.3
2034	18.2	18.7	18.2	17.1	18.3	15.7	11.4	11.5	17.5	18.0	17.6	16.7	18.1	16.2	13.4	16.3
2039	18.3	18.3	18.6	17.2	17.3	16.8	11.2	11.1	17.6	17.6	18.0	16.9	17.1	17.4	13.1	15.9
2044	18.6	18.1	18.5	17.9	16.5	16.7	12.6	10.0	17.8	17.4	17.9	17.5	16.4	17.3	14.7	14.3
2049	18.7	18.2	18.1	18.3	16.7	15.8	13.4	9.7	17.9	17.5	17.5	17.9	16.6	16.4	15.7	13.9
2054	18.6	18.5	17.8	18.1	17.3	15.1	13.4	10.1	17.8	17.7	17.3	17.7	17.2	15.7	15.7	14.4
2059	18.4	18.6	18.0	17.7	17.7	15.3	12.6	10.7	17.7	17.8	17.4	17.4	17.6	15.8	14.8	15.2
2064	18.4	18.5	18.2	17.5	17.5	15.8	12.1	10.9	17.6	17.8	17.6	17.1	17.4	16.4	14.2	15.4
2069	18.5	18.3	18.3	17.6	17.2	16.2	12.2	10.6	17.7	17.6	17.7	17.3	17.0	16.8	14.3	15.0
2074	18.6	18.3	18.2	17.9	17.0	16.1	12.7	10.2	17.8	17.6	17.7	17.5	16.8	16.6	14.9	14.6
2079	18.5	18.4	18.1	18.0	17.1	15.7	12.9	10.2	17.8	17.6	17.5	17.6	16.9	16.3	15.2	14.5
2084	18.5	18.4	18.0	17.9	17.3	15.5	12.8	10.4	17.7	17.7	17.5	17.5	17.2	16.1	15.0	14.7

Note. Computations were carried out through the year 2119, but results are shown only to the point at which values appeared to stabilize. Units = one million persons.

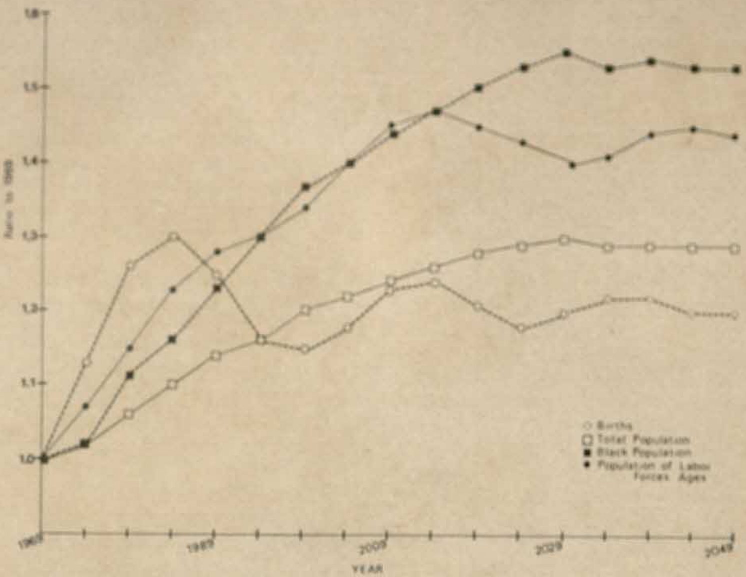


FIGURE 1. Estimated demographic changes at five-year intervals assuming that a replacement birthrate was attained in 1969.

Table 1. The trends shown in Table 1 appear to be consistent with other estimates of a gradual transition to a stable population in the United States (Ehrlich & Ehrlich, 1972). The total population continued to increase for some 60 to 65 years after attainment of a replacement birthrate, with a final population about 30% larger than the base population. The proportion of children in the population decreased; the proportion of aged in the population increased; and a series of waves or peaks moved through the age distribution (Frejka, 1968). The first wave represented persons born in the baby-boom after World War II; the second wave represented their children, etc. As a consequence of these waves, the age distribution would not reach a steady state for a number of years after the population size stabilized.

Some other demographic consequences of a replacement birthrate are presented in Figure 1. Data points in this figure represented ratios of the number of people in a given category in a given year to the number of people in the same category in 1969. The birthrate stabilized in some 70 to 75 years, after considerable fluctuation because of variation in the number of women in the childbearing years. The relative increase in population size was greater than the relative increase in births, largely because the use of the death rates of the white population for the death rates of the entire population of the United States produced an increase in overall life expectancy. For the same reason, the relative increase in the population of blacks, or Chicanos, or American Indians, etc., was considerably greater than the relative increase in the total population. These results reduce the plausibility of the charge of "genocide," so long as the search for population stability was combined with a search for social justice. (However, relatively little change in the proportion of blacks in the total population occurs.) Finally, the labor force (i.e., persons in the age range 20–65) grew both more and faster in size than the total population.

Changes in mental abilities. Estimates by Schaie and Strother (1968) and Schaie (1965, 1970) of the longitudinal relationship between age and *T* scores on Thurstone's (1958) SRA Primary Mental Abilities Tests were used for the present study. Changes between the ages of 21 and 70 were estimated from results of Schaie and Strother's youngest cohort. Changes after age 70 were estimated by assuming that the percentage of

TABLE 2

ESTIMATED RELATIONSHIP BETWEEN AGE AND INTELLECTUAL ABILITIES

Age	Intellectual ability			
	Verbal meaning	Reasoning	Number	Word fluency
0–4	11.5	12.5	10.8	11.5
5–9	23.8	25.7	22.4	23.7
10–14	37.2	40.2	34.9	37.1
15–19	46.5	50.3	43.7	46.4
20–24	49.9	54.4	45.9	50.5
25–29	51.9	55.5	49.8	50.8
30–34	52.0	56.1	53.2	49.7
35–39	51.1	56.0	55.1	46.0
40–44	50.8	55.7	57.3	45.4
45–49	49.7	53.4	58.0	44.3
50–54	48.2	51.4	57.3	40.4
55–59	46.5	49.3	55.3	35.7
60–64	45.6	46.9	52.3	30.9
65–69	44.7	45.7	51.7	27.7
70–74	43.8	44.5	50.9	24.8
75–79	42.9	43.4	50.1	22.3
80–84	42.1	42.3	49.4	20.0
85–89	41.3	41.2	48.6	17.9
90–94	40.4	40.2	47.9	16.0
95–99	39.6	39.1	47.1	14.4
100–104	38.9	38.1	46.4	12.9

Note. Entries are based on T scores with $\bar{X} = 50$ and $SD = 10$.

change between adjacent age intervals are the same as the percentage of change between Schaie and Strother's last two age intervals. Changes prior to age 21 were computed on the assumption that Bayley's (1955) chart of intellectual development adequately describes the development of all traits.[4] The final relationships between age and abilities produced by these assumptions are summarized in Table 2. (The Space test is omitted because Schaie and Strother's estimates for this test appeared unstable.) These relationships conform to current psychological notions about the relationship between age and mental ability in that childhood and adolescence are characterized by steady growth; ability reaches a peak during early maturity; there is a relatively slow decline during the middle years, and a more rapid decline in old age. The trends shown in Table 2 also appear generally consistent with Schaie's (1970) latest estimates of maturational changes in those abilities.

The figures shown in Tables 1 and 2 were combined to yield estimated average ability scores in various years for the total population and for the labor force. Table 3 summarizes this analysis. It is apparent that only minimal changes in average ability would occur solely from a change to a replacement birthrate, and therefore no suggestion that any special problems for society would result from changes in average ability. There was no evidence, for example, that the larger proportion of older people would produce a noticeable decline in average ability.

An obvious question is the effect of cohort differences on these conclusions. Schaie's (1970) latest analyses showed that, in general, older cohorts have lower average ability scores with some variation among cohorts. If cohort differences were taken into account, one would expect slightly lower population averages in the early years after attainment of a replacement birthrate, but no important deviations from the pattern shown in Table 3. One would also expect values to converge quickly to the values shown in Table 3 inasmuch as persons in the same age ranges as Schaie and Strother's cohorts represent a rapidly decreasing proportion of the total population. (The mean year of birth for Schaie and Strother's youngest cohort was 1940. People in the same age ranges as their cohorts represented only about 55% of the population in 1969 and by the year 2009 will represent only about 10% of the population.) These expectations were confirmed by computing population averages on verbal meaning, taking cohort differences into account. It appears, therefore, that Table 3 provides a reasonable estimate of the combined effects of changes in the age distribution and maturational changes in these mental abilities.

Changes in personality patterns. One of the few studies relating age to personality, which is relevant to this study, appeared to be Bendig's (1960) investigation (based on men in Pittsburgh) of age differences in the Guilford-Zimmerman Temperament Survey. Al-

---

[4] Thurstone (1955), as cited by Bloom (1964), proposes a slightly different course of early development for these abilities but not in a form convenient for deriving estimates like those used in this study.

TABLE 3

ESTIMATED MEAN INTELLECTUAL ABILITY SCORES FOR POPULATION IN VARIOUS YEARS

Year	Intellectual ability scores							
	Population				Labor force			
	Verbal meaning	Reasoning	Number	Word fluency	Verbal meaning	Reasoning	Number	Word fluency
1969	41.6	44.5	43.4	37.0	49.8	53.5	53.4	44.7
1974	42.6	45.6	44.5	38.0	49.8	53.6	53.2	44.9
1979	43.0	46.0	44.9	38.3	49.9	53.7	53.0	45.1
1984	42.9	46.0	44.9	38.1	50.0	53.8	53.0	45.3
1989	42.8	45.8	44.9	37.9	50.1	53.9	53.3	45.3
1994	42.9	45.9	45.3	37.9	50.1	53.9	53.8	45.1
1999	43.2	46.3	45.7	38.1	49.9	53.8	54.0	44.8
2004	43.4	46.5	46.1	38.2	49.8	53.5	53.9	44.4
2009	43.4	46.5	46.2	38.1	49.7	53.3	53.8	44.2
2014	43.3	46.3	46.1	37.7	49.6	53.3	53.7	44.0
2019	43.2	46.1	46.0	37.4	49.6	53.3	53.7	43.9
2024	43.2	46.1	46.1	37.2	49.7	53.4	53.7	44.0
2029	43.3	46.2	46.2	37.2	49.8	53.5	53.8	44.3
2034	43.3	46.2	46.3	37.3	49.8	53.5	53.8	44.4
2039	43.3	46.2	46.2	37.3	49.7	53.4	53.7	44.2
2044	43.3	46.2	46.2	37.3	49.7	53.3	53.7	44.0
2049	43.3	46.2	46.1	37.2	49.7	53.3	53.7	44.0
2054	43.3	46.2	46.2	37.2	49.7	53.4	53.7	44.1

though obvious questions arise about the extent to which these results were representative of the total population of men of the United States, some data are better than none, and accordingly, age group means obtained by Bendig (on the five scales showing significant age differences) were combined with the age distributions shown in Table 1 to yield estimated population means at 15-year intervals. Results are summarized in Table 4. As was apparent from Bendig's means, again there was virtually no change in average scores for the population as a consequence of a replacement birthrate. In other words, there was no evidence that the typical personality pattern for the adult male population would change very much.

Another way to estimate changes in personality patterns is through changes in the incidence of various psychiatric disorders. Accordingly, Jaco's (1960) estimates (based on Texans of both sexes combined) of the age specific incidence of functional, organic, and old age psychoses were used to estimate incidence in the total population (i.e., cases per 100,000 people) in various years after attainment of a replacement birthrate. Table 5 summarizes the results. There was some increase in incidence of all three types of psychosis, as well as some differential change in

TABLE 4

ESTIMATED MEAN GUILFORD–ZIMMERMAN SCORES FOR MALES AGE 20–59 IN VARIOUS YEARS

Year	Males				
	Restraint	Ascendance	Sociability	Masculinity	Personal relations
1969	19.98	18.41	22.42	20.04	22.42
1984	19.89	18.58	22.59	20.13	22.45
1999	20.08	18.32	22.35	19.97	22.52
2014	20.04	18.30	22.30	19.97	22.49
2029	20.06	18.33	22.35	19.98	22.49
2044	20.05	18.31	22.32	19.97	22.50
2059	20.06	18.31	22.33	19.97	22.49

Note. Entries are estimated raw score means. Only scales showing significant age differences in Bendig's (1960) study are included.

TABLE 5

ESTIMATED TRENDS IN INCIDENCE OF FUNCTIONAL, ORGANIC, AND OLD AGE PSYCHOSES

Year	Incidence per 100,000 persons		
	Functional	Organic	Old age
1969	49.2	9.3	14.7
1974	51.4	9.7	14.9
1979	53.3	9.9	15.3
1984	54.5	10.1	15.7
1989	54.8	10.2	16.1
1994	55.0	10.4	16.4
1999	55.4	10.6	16.2
2004	56.3	10.9	16.1
2009	56.9	11.3	16.6
2014	56.7	11.4	17.9
2019	56.1	11.3	19.3
2024	55.3	11.2	21.0
2029	54.6	11.2	22.4
2034	54.8	11.2	22.3
2039	55.4	11.3	21.2
2044	55.7	11.3	20.6
2049	55.5	11.3	20.8
2054	55.2	11.2	21.4

that the percentage of increase was somewhat greater for old age psychoses than for the other psychoses. Nevertheless, the overall pattern varied very little as a consequence of the changes in the age distributions.

Thus, two different procedures for examining population personality change yielded results that were quite consistent with each other and with the results for intellectual ability presented earlier. A change to a replacement birthrate should produce no striking changes in population personality characteristics that would pose serious problems for society.

Changes in life transitions. In recent years, largely through Erickson's (1950) influence, psychological changes through the life span have increasingly been identified with important life transitions. A corollary of this approach is that the overall psychological characteristics of a society could be determined, in part, by the relative incidence of various life transitions in that society. Accordingly, the last analysis involved changes in the incidence of three important life transitions to which it appeared reasonable to assign a specific age. These were: (*a*) entering school (age = 6), (*b*) leaving the parental home (age = 18), and (*c*) retiring (age = 65). Changes in the incidence of these transitions for both sexes combined are summarized in Figure 2.

Results in Figure 2 show (*a*) the influence of the waves or peaks that pass through the age distribution, (*b*) an overall decrease in the incidence of entering school[5] and leaving home (the absolute number of persons going through these transitions also would decrease), and (*c*) an overall increase in the incidence of retiring. Similar trends are expected for life transitions that are age related but that cannot reasonably be assigned to specific ages. For example, there would be a relative decrease in the incidence of first marriages and a relative increase of persons facing the death of a spouse.

The overall effect of the trends shown in Figure 2 was to bring the curves closer together but not to change the overall pattern drastically. More individuals still would enter school and leave home each year than those that would retire, even after population stabilizes. This finding suggests that changes in relative incidence of life transitions are insufficient to produce drastic changes in the overall psychological characteristics of the United States society. Therefore, it again appears that no major social dislocations would occur in the United States solely as a consequence of a change to a replacement birthrate. Indeed, some problems could be eliminated, in that society is no longer faced with wide fluctuation over relatively brief periods in such transitions as entering school.[6]

CONCLUSIONS

While more comprehensive replications with better data about the relationship of age

---

[5] Birth data in Figure 1 involve total births over five-year intervals, while entering school data in Figure 2 involve incidence in a single year that is not at the midpoint of its age interval. Any apparent discrepancies between these figures can be attributed to these procedural differences.

[6] It should be emphasized that for about the next fifty years the fluctuations shown in Figure 2 will occur whether or not a change to a replacement birthrate occurs.

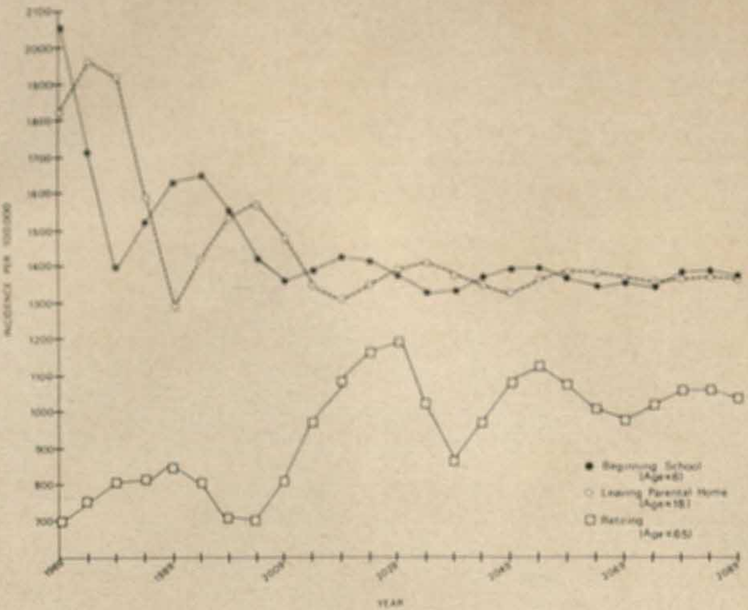


FIGURE 2. Estimated incidence at five-year intervals of three important life transitions.

to psychological characteristics are needed, the results of this study suggest that changes in such characteristics do not appear to pose any serious objection to a replacement birthrate in the United States. This is not to say that possible psychosocial consequences can be entirely ignored in striving to attain and maintain a replacement birthrate. A stable population, especially if combined with a steady-state economy (Daly, 1971), could produce major changes in the structure of society, and these changes could involve great psychological changes. Specifically, a stable population and economy probably involve much more stability in the overall society. This might produce a much less positive attitude toward change and toward achievement. Other consequences might be a much less youth-oriented society, with much more emphasis on age and experience as positive values. All of these changes could produce large psychosocial changes in the United States population. These changes, however, would result from a different way of organizing society rather than being a necessary consequence of the change to a replacement birthrate itself. Therefore, the possibility of such changes also does not appear to pose a serious objection to a replacement birthrate.

Since population problems, both practical and scientific, confront all of mankind, not just the population of the United States, similar investigations in other nations could be helpful, especially in nations whose population age distributions are quite different from the United States distribution.

REFERENCES

AMERICAN PSYCHOLOGICAL ASSOCIATION TASK FORCE. Report of the task force on psychology, family planning, and population policy. *American Psychologist*, 1972, **27**, 1100–1105.

BAYLEY, N. On the growth of intelligence. *American Psychologist*, 1955, **10**, 805–818.

BENDIG, A. W. Age differences in the interscale factor structure of the Guilford–Zimmerman Temperament Survey. *Journal of Consulting Psychology*, 1960, **24**, 134–138.

BLOOM, B. S. *Stability and change in human characteristics*. New York: Wiley, 1964.

COALE, A. J. *The growth and structure of human population: A mathematical investigation*. Princeton, N.J.: Princeton University Press, 1972.

DALY, H. E. Toward a stationary-state economy. In J. Harte & R. H. Socolow (Eds.), *Patient earth*. New York: Holt, Rinehart & Winston, 1971.

Ehrlich, P. R., & Ehrlich, A. A. *Population, resources, environment.* (2nd ed.) New York: Freeman, 1972.

Erickson, E. H. *Childhood and society.* New York: Norton, 1950.

Fawcett, J. T. *Psychological perspectives on population.* New York: Basic Books, 1973.

Frejka, T. Reflections on the demographic conditions needed to establish a U.S. stationary population growth. *Population Studies,* 1968, **22**, 379–397.

Hutchinson, J. (Ed.) *Population and food supply.* Cambridge, England: Cambridge University Press, 1969.

Jaco, E. G. *The social epidemiology of mental disease.* New York: Russell Sage Foundation, 1960.

Kirk, D. Review of *The limits to growth. Contemporary Psychology,* 1973, **18**, 1–3.

Looft, W. R. The psychology of more. *American Psychologist,* 1971, **26**, 561–565.

Meadows, D. H., Meadows, D. C., Randers, J., & Behrens, W. W., III. *The limits to growth.* New York: Universe Books, 1972.

NAS–NRC Committee on Resources and Man. *Resources and man.* San Francisco, Calif.: Freeman, 1969.

Pohlman, E. *The psychology of birth planning.* Cambridge, Mass.: Schenkman, 1969.

Schaie, K. W. A general model for the study of developmental problems. *Psychological Bulletin,* 1965, **64**, 92–107.

Schaie, K. W. A reinterpretation of age-related changes in cognitive structure and functioning. In L. R. Goulet & P. B. Baltes (Eds.), *Life-span developmental psychology: Research and theory.* New York: Academic Press, 1970.

Schaie, K. W., & Strother, C. R. The effect of time and cohort differences on the interpretation of age changes in cognitive behavior. *Multivariate Behavioral Research,* 1968, **3**, 259–293.

Thurstone, L. L. *The differential growth of mental abilities.* (Psychometrics Lab. Rep. 14) Chapel Hill: University of North Carolina Press, 1965.

Thurstone, T. G. *Manual for the SRA Primary Mental Abilities Tests.* Chicago: Science Research Associates, 1958.

(Received November 6, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 9–14

INFLUENCE OF SEX ROLE STEREOTYPES ON PERSONNEL DECISIONS

BENSON ROSEN[1] AND THOMAS H. JERDEE

Graduate School of Business Administration, University of North Carolina

An in-basket exercise was used to investigate the influence of sex role stereotypes on the personnel decisions of 95 bank supervisors. The design consisted of four separate experiments (in-basket items) in which an employee's sex and other situational attributes were manipulated. Results confirmed the hypothesis that male administrators tend to discriminate against female employees in personnel decisions involving promotion, development, and supervision. Discrimination against male employees was found in personnel decisions involving competing role demands stemming from family circumstances. Features of the decision-making process which tend to heighten the influence of sex role stereotypes on personnel decisions are discussed.

It is frequently alleged that male administrators view females as equipped to do the organizational housekeeping but as deficient in the toughness, stability, judgment, and dedication required for success in managerial and other traditionally male roles. To protect both the organization and the "vulnerable" female employee, male administrators allegedly resort to a pattern of exclusion in selection, promotion, and development which bars women from the more challenging organizational roles or places them at a disadvantage when they do achieve these roles.

In some ways, however, the commonly alleged sex role stereotypes may tend to work in favor of females and against males. Since men are expected to pursue their careers with single-minded dedication, any intrusion of family or other personal considerations may be viewed more unfavorably for them than for women. In cases involving time off or other similar indulgences for family and personal reasons, management may deal more harshly with males than with females, and young male managerial aspirants may be reluctant to make such requests for fear of jeopardizing their careers.

In spite of all the allegations and circumstantial evidence, the actual behavioral effects of sex role stereotyping have only recently been subject to systematic investigation (see, for example, studies by Goldberg, 1968; Shaw, 1972; Rosen & Jerdee, 1973). While some studies have supported the general notion of the biasing effects of sex role stereotypes, they cover a rather narrow range of discriminatory behavior and they do not deal at all with those situations in which sex role stereotypes may work to the disadvantage of males. Accordingly, it is desirable to direct research at both types of effects, those forms of exclusion and favoritism favoring males and those forms favoring females. On the basis of the commonly alleged stereotypes for males and females, we hypothesized (*a*) that male administrators would tend to discriminate against females in important decisions involving promotion, development, and supervision and (*b*) that they would tend to discriminate against males in decisions involving competing role demands stemming from family or other personal circumstances. These hypotheses were tested in an experiment involving a group of male managers from the banking industry.

METHOD

Subjects

The subjects were 95 male bank supervisors attending a management institute at the University of North Carolina during the summer of 1972. The institute was open to both sexes, but since only 13 women attended, it was necessary to restrict the study to males.

---

[1] Requests for reprints should be sent to Benson Rosen, Graduate School of Business Administration, University of North Carolina, Chapel Hill, North Carolina 27514.

Procedure

Experimental materials were embedded in an in-basket exercise.[2] Subjects were asked to assume the role of the personnel director of a multibranch bank and to react to a series of items in memorandum or letter form. Four of these in-basket items covered different types of personnel problems and were written in two or more versions so as to manipulate the sex of one or more of the characters and certain other attributes of the characters and the situation.

For each item, subjects indicated on fixed-response scales their decisions and the extent to which they would find certain reactions to the case acceptable. Each subject received only one version of each of the experimental items and was unaware that other versions existed. From the subjects' perspective, all were participating in an identical in-basket exercise dealing with personnel problems in the banking industry. The sex variable was completely unobtrusive. Since each item dealt with a different type of problem and involved different dependent variables (managerial decisions), the design can best be conceptualized as a series of four separate experiments with different experimental stimuli and different dependent variables in each and with completely random assignment of treatments to subjects within each experiment.

Three of the in-basket items involved administrative decisions in which sex role stereotypes were hypothesized to work to the disadvantage of females. These items dealt with promotion, development, and leadership style. These three items and the experimental variables of interest were as follows:

1. *Promotion to branch manager: Sex of candidate and complexity of job.* This item was in the form of a memorandum requesting a decision on the promotion of an employee to the position of branch manager. The memorandum was written in four versions so as to manipulate the variables of sex of the candidate and the nature of the manager's job, which was described as either complex or routine. We expected a greater tendency to promote the male, especially when the job was more complex.

The major dependent variables were (*a*) a recommendation either to promote the candidate or to hold the candidate's file and interview additional candidates, (*b*) a rating of the candidate on potential for customer relations, and (*c*) a rating of the candidate on potential for employee relations. Subjects were asked to make their evaluations on 6-point scales ranging from extremely unfavorable (1) to extremely favorable (6).

2. *Development: Choice of a male or female candidate to attend a conference.* This item was in the form of a memorandum asking the personnel director to decide which of two persons on his staff should attend a personnel conference. In one version of the memorandum the choice was between an older, unpromotable female and a younger, highly qualified male. In the other version, the choice was between an older, unpromotable male and a younger, highly qualified female. We expected that the male would be chosen more frequently than the female.

3. *Solution of a supervisory problem: Sex of Supervisor × Sex of Subordinate × Nature of the Problem.* This item consisted of a lengthy report from a bank supervisor describing a problem with one of his (her) subordinates and requesting the termination or at least the transfer of the subordinate. We expected that the subjects would be more inclined to accept the supervisor's judgment and proceed with the termination or transfer when the supervisor was a male than when the supervisor was a female. We also sought to examine how the sex of the subordinate and the nature of the problem (poor performance vs. offensive personality) would affect the subjects' willingness to terminate or transfer the subordinates. Accordingly, eight versions of this case were used, forming a 2 × 2 × 2 factorial design.

The major dependent variables for this item were subjects' ratings of the following administrative actions: (*a*) arrange for a termination of the subordinate; (*b*) arrange for transfer of the subordinate; (*c*) suggest that the supervisor try again to solve the problem; or (*d*) arrange for an interview with both the supervisor and the subordinate.

4. *Approval of request for a leave of absence: Sex of employee.* The final item involved an administrative decision in which sex role stereotypes were hypothesized to lead to more indulgent or deferent behavior toward females than toward males. This item was in the form of a request from an accountant for time off to take care of children. In Version A, the request was made by a female accountant who claimed that it was impossible for her husband to assume these responsibilities. Version B involved the identical situation except that the accountant was a male whose wife's career obligations prevented her from assuming the child care responsibilities. We expected that subjects would be willing to grant the leave to the female accountant, since it would be more in keeping with the stereotype regarding the primacy of a woman's family obligations.

Subjects were asked to rate the appropriateness of the request and the desirability of the following administrative actions: (*a*) give the employee a leave of absence without salary; (*b*) do not allow the employee to take a leave of absence; and (*c*) allow the employee to take a leave of absence with salary.

RESULTS

Our first major hypothesis stated that male administrators would tend to discriminate against females in important decisions involving promotion, development, and supervisory practice. Results for the in-basket items relevant to this hypothesis will now be considered.

---

[2] See Fredricksen (1968) for an example of the use of the in-basket technique as a research task.

TABLE 1

PROMOTION CASE: ADMINISTRATIVE ACTIONS FOR FEMALE AND MALE CANDIDATES

Administrative action	Female candidate			Male candidate			Total
	Simple job ($n = 24$)	Complex job ($n = 25$)	Both jobs	Simple job ($n = 24$)	Complex job ($n = 20$)	Both jobs	
Promote[a]	14	5	19	21	11	32	51
Hold file[a]	10	20	30	3	9	12	42
Potential for customer relations[b]	4.52	4.01	4.25	4.95	4.51	4.73	4.47
Potential for employee relations[b]	3.99	4.39	4.19	4.75	4.66	4.70	4.43

[a] Number of subjects.
[b] Mean rating.

Promotion to Branch Manager

Subjects' promotion decisions and mean ratings of the candidates are shown in Table 1. From Table 1 it can be seen that subjects were more willing to promote a male candidate than a female candidate. The male candidate was recommended for promotion by 32 out of 44 subjects, while the female candidate was recommended for promotion by only 19 out of 49 subjects ($\chi^2 = 6.53$, $df = 1$, $p < .05$). As might be expected, subjects were more willing to promote a candidate to a simple job than to a complex job ($\chi^2 = 13.09$, $df = 1$, $p < .01$). The interaction of job complexity by sex of candidate was not significant. It appears that subjects discriminated against women in their promotion recommendations and that the bias against promoting women to managerial positions held across the two conditions of job complexity.

Table 1 also shows mean ratings of each candidate's potential in the experimental conditions. Ratings of potential for customer relations were influenced by both sex and job type. Mean rating for the male candidate was 4.73 compared to a mean rating of 4.25 for the female candidate ($F = 4.76$, $df = 1/80$, $p < .05$). The mean rating for candidates for the simple job was 4.72 compared to a mean of 4.24 for the complex job ($F = 4.93$, $df = 1/80$, $p < .05$). The sex of the candidate by job complexity interaction was not significant.

Finally, Table 1 shows the mean rating of the candidate's potential for employee relations in each of the experimental conditions. The sex of the candidate variable was again significant. The mean rating on potential for employee relations for the male candidate was 4.70 compared to 4.19 for the female candidate ($F = 4.57$, $df = 1/80$, $p < .05$). No significant effects were obtained for the job complexity variable or for the complexity by sex or candidate interaction. These results clearly demonstrate discrimination against females in promotion decisions.

Choosing an Employee to Attend a Professional Training Conference

Table 2 shows subjects' choices of an employee to participate in the professional training conference in the two experimental conditions.

When the younger, promotable employee was male, 76% of the subjects selected him in preference to the older unpromotable female. On the other hand, when the younger, promotable employee was female, only 56% selected her in preference to the older, unpromtotable male ($z = 2.04$, $p < .05$). It appears that a highly promotable male employee is strongly preferred to a female employee with less potential; however, a highly promotable female is preferred only slightly more frequently than an unpromotable male.

Results for this in-basket item add further support to the hypothesis that women are victims of discrimination in important personnel decisions involving career progress.

Solution of a Supervisory Problem

Table 3 shows the mean rating for each of the administrative actions under each of the eight experimental conditions. Ratings of

TABLE 2

SELECTION OF EMPLOYEE TO ATTEND CONFERENCE

Male		Female	
No. subjects	Proportion of row total	No. subjects	Proportion of row total
Version A: Qualified, younger male versus unpromotable, older female			
37	.76	12	.25
Version B: Qualified younger female versus unpromotable, older male			
20	.44	25	.56
Total			
57	.61	37	.39

termination and of transfer are of particular interest and will be considered together since these are mutually exclusive actions with negatively correlated ratings.

Main effects were not significant but there were significant interaction effects (Sex of Supervisor × Nature of Problem) for both termination ($F = 5.30$, $df = 1/69$, $p < .05$) and transfer ($F = 4.75$, $df = 1/69$, $p < .05$). For *performance* problems, the differential effects of the supervisor's sex are quite clear. When the requesting supervisor was male, termination of the subordinate was rated high ($\bar{X} = 4.3$) and transfer low ($\bar{X} = 2.9$); when the requesting supervisor was female, termination of the subordinate was rated low ($\bar{X} = 2.9$) and transfer high ($\bar{X} = 4.2$). One speculative interpretation of these findings is that subjects had greater confidence in the ability of a male supervisor than a female supervisor to appraise the seriousness of a performance problem. Accordingly, subjects would be more willing to accept the male supervisor's recommendation for termination.

For *personality* problems, the differential effects of sex are not as clear. When the requesting supervisor was male, termination of the subordinate was rated low ($\bar{X} = 2.9$) and transfer high ($\bar{X} = 4.1$); when the requesting supervisor was female, termination and transfer rated equally ($\bar{X} = 3.5$). The differential effects of sex here are difficult to interpret.

The results for the three in-basket items reported thus far strongly support the hypothesis that sex role stereotypes lead to discrimination in personnel decisions involving promotion, development, and supervision. We now turn to results relevant to our second hypothesis. Here we predicted that males who deviated from the sex role stereotypes that depicted men as dedicated to their work above family and personal considerations would be dealt with more harshly than females in similar situations. The final in-basket item is concerned with this hypothesis.

Approving a Leave of Absence

Mean ratings for the leave request and alternative administrative actions are shown in Table 4. Subjects rated the request for a

TABLE 3

SUPERVISOR–SUBORDINATE CONFLICT: MEAN RATING OF ADMINISTRATIVE ACTIONS FOR EACH EXPERIMENTAL CONDITION

Managerial action	Male supervisor				Female supervisor				Row $\bar{X}$
	Performance		Personality		Performance		Personality		
	Male ($n = 8$)	Female ($n = 6$)	Male ($n = 11$)	Female ($n = 7$)	Male ($n = 10$)	Female ($n = 10$)	Male ($n = 9$)	Female ($n = 8$)	
Arrange a termination	4.25	4.42	3.38	2.62	3.00	3.00	4.00	3.10	3.42
Arrange a transfer	3.00	2.85	3.45	4.90	4.00	4.20	3.10	4.00	3.74
Suggest that supervisor try to resolve problem	2.00	2.16	2.45	2.00	1.90	2.40	1.88	2.00	2.11
Arrange for interview with both parties	5.40	4.42	4.47	4.62	5.18	5.00	4.41	5.60	4.87

leave to care for small children as significantly less appropriate when it came from a male accountant than when it came from a female accountant. In addition, subjects rated it significantly more appropriate for a female than a male to take a leave of absence without salary. Here we have a situation where the male sex role stereotype of complete dedication to the job appears to be influencing subjects' ratings. Subjects clearly perceive it more inappropriate for males than for females to take time off because of family obligations.

DISCUSSION

Our findings support current allegations regarding discrimination against women in a variety of organizational situations. The pervasiveness of the discrimination was demonstrated by the finding of bias against women in personnel decisions related to promotion, development, and supervision.

In view of the unobtrusiveness of the manipulations in the experiment (subjects were unaware of other versions of each in-basket item), the results demonstrating the effect of sex role stereotypes seem impressive. Subjects' responses in the simulated decision-making situation provided by the in-basket exercise were similar in a very important way to comparable administrative decisions in real life; that is, each administrative decision was made for only one person (sometimes a male and sometimes a female). It might be reasonable to assume that when managers are confronted with both a male and a female employee for either a new position, promotion, advanced training, or supervisory assistance, they will be more likely to act in a fair and impartial way. In such instances, the potential for discrimination is quite salient and subjects might tend to guard against favoritism. On the other hand, in many organizational settings such administrative actions occur as single events, as was the case in our experimental situation. In these situations, where the possibility of discrimination is not an obvious factor in the decision process, the influence of sex role stereotypes seems to have a definite impact.

TABLE 4

REQUEST FOR A LEAVE OF ABSENCE: MEAN RATINGS OF APPROPRIATENESS OF REQUEST AND ADMINISTRATIVE ACTIONS

Item	Male accountant ($n = 43$)	Female accountant ($n = 46$)	t	p
Appropriateness of request	1.95	2.64	2.07	.05
Give leave without salary	2.88	4.06	3.12	.01
Do *not* allow accountant to take leave of absence	3.88	3.44	1.00	*ns*
Give leave with salary	1.69	1.43	1.07	*ns*

Our data reveal a second feature of the decision situation which tends to foster biased decisions. Indications of discrimination were found with regard to administrative actions that were not unanimously perceived as clearly appropriate or clearly inappropriate. For example, discrimination was quite strong in the case where subjects were asked to judge the "potential for customer and employee relations" of a candidate for promotion to branch bank manager and also in the case where subjects were asked to determine the desirability of granting a leave of absence without pay to a bank accountant. In situations where the desirability or appropriateness of a particular administrative action is ambiguous and only partial information is available to the decision maker, subjects appear to fall back on preconceived attitudes (sex role stereotypes) to arrive at their ultimate decisions. In a sense, subjects may be providing more information for themselves by filling in either their own prejudices or some widely held societal expectations.

Our findings also demonstrate what may become an increasingly serious role conflict for male employees: the conflict between career and family responsibilities. While most of the discussion of sex role stereotypes has centered on female role conflicts between career and family, it is likely that male employees may soon experience a similar dilemma. In spite of the obstacles depicted in the Results section, more women are moving into managerial and professional careers. Accordingly, it seems reasonable to assume that the husbands of career women will be expected to share many of the family responsibilities which previously had been considered the female's province. This means that changes in attitudes may be necessary with regard to the appropriate roles for both males and females in professional and private life.

One way to broaden managerial attitudes regarding male and female roles is to demonstrate to managers the restrictive effects of current stereotypes. The in-basket as used in this study offers some promise for this purpose. By collecting data early in the management training program, we were able to feed back our results to participants in a later session. This provided an opportunity to dramatize to our subjects the many subtle ways in which their own decisions were influenced by sex role stereotypes.

REFERENCES

FREDRIKSEN, N. *Organizational climates and administrative performance.* Princeton, N.J.: Educational Testing Service, 1968.

GOLDBERG, P. Are women prejudiced against women. *Transaction*, 1968, **5**, April, 28–30.

ROSEN, B., & JERDEE, T. H. The influence of sex-role stereotypes on evaluations of male and female supervisory behavior. *Journal of Applied Psychology*, 1973, **57**, 44–48.

SHAW, E. A. Differential impact of negative stereotypes in employee selection. *Personnel Psychology*, 1972, **25**, 333–338.

(Received August 28, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 15–23

THE LIKERT ORGANIZATIONAL PROFILE: METHODOLOGICAL ANALYSIS AND TEST OF SYSTEM 4 THEORY IN BRAZIL[1]

D. ANTHONY BUTTERFIELD
School of Business Administration
University of Massachusetts

GEORGE F. FARRIS[2]
Alfred P. Sloan School of Management
Massachusetts Institute of Technology

A 20-item Likert Organizational Profile was administered twice to 256 employees in 13 Brazilian development banks. Actual and ideal bank profiles were similar to those found in the United States and elsewhere: employees want participative group management methods but say that their organization uses autocratic or consultative methods. Factor analyses did not yield the six dimensions predicted by Likert's theory; factors were only partially consistent over time and for different hierarchical levels. Retrospective scores were quite accurate and were equally so for the 6-, 12-, and 18-month time periods since the original administration. Test–retest reliability of the Likert Organizational Profile as a whole was moderate. Bank Likert Organizational Profile scores were unrelated to objective measures of organizational effectiveness, but were positively related to employee satisfaction. The Likert Organizational Profile appears useful for current and retrospective organization studies, but the theory of management systems measured by it was only partially supported.

The work of Likert and his associates at the Institute for Social Research represents a major contribution to modern organization theory. Likert's two books (1961, 1967) summarize much of the work of the institute and present his own theory of management. The theory postulates four systems of management: (*a*) authoritative, (*b*) benevolent authoritative, (*c*) consultative, and (*d*) participative group. The key organizational characteristics of the different systems are defined in terms of leadership, motivation, communication, decision making, goal setting, and control processes. System 4 (participative group) is seen as doing the best job of utilizing the organization's human assets, thereby producing greater employee motivation and more favorable attitudes as well as more effective organizational performance. A considerable amount of empirical research has been carried out on the theory (e.g., Likert, 1961, 1967; Marrow, Bowers, & Seashore, 1967); however, most of the research was done in organizations in the United States.

More recently, the theory has been tested in Yugoslavia and Japan using various versions of a "profile of organizational characteristics" (Kavčič, Rus, & Tannenbaum, 1971; Likert, 1969). The profile is a questionnaire that describes an organization on the dimensions of Likert's theory and has come to be known as the Likert Organizational Profile. The Likert Organizational Profile has also been used in the United States, frequently in studies of change (e.g., Blumberg & Wiener, 1971; Golembiewski & Carrigan, 1970a, 1970b; Marrow et al., 1967).

Despite this use of the Likert Organizational Profile in organization studies, it has not been subject to much analysis as a measuring instrument. Little evidence has been presented as to the Likert Organizational Profile's reliability or factor structure. The purpose of the present paper is to present a methodological analysis of the Likert Organizational Profile and to examine the suitability of using the Likert Organizational Profile to study organizations in a country outside the

---

[1] This study was supported by a grant from the Ford Foundation. The authors wish to thank members of the Grupo de Trabalho Misto for their help and Rensis Likert for his permission to use the Likert Organizational Profile. Special acknowledgement goes to Cecilio Berndsen and Eldon Senner for their assistance in gathering and processing the data.

[2] Requests for reprints should be sent to George F. Farris, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

United States—Brazil in the present case. We shall also test support for Likert's theory in Brazil.

To meet these objectives we shall (*a*) present profiles of Brazilian organizations as seen by their members in terms of Likert's four management systems; (*b*) discuss factor analyses of the Likert Organizational Profile; (*c*) present a longitudinal analysis of the profile; and (*d*) examine relationships between Likert Organizational Profile scores and organizational effectiveness and employee satisfaction.

METHOD

Instrument

The Likert Organizational Profile used in the present study was Form S, distributed by the Foundation for Research on Human Behavior (1967). Two items (Number 2, subordinates' confidence in superiors, and Number 12, subordinates' knowledge of organizational policies) were added at the request of the Brazilians, making a total of 20 items on the profile. Questions were translated from English into Portuguese by bilingual Brazilians. These translations were then checked and revised by the authors in collaboration with the translators until all were satisfied that the translations were as accurate as possible. An independent retranslation into English indicated that the Portuguese and English versions were essentially identical.

Likert Organizational Profile questions are answered on a 20-point scale divided into four sections, each section representing one of the four management systems described by Likert's theory. Each question has a different set of four descriptive alternatives corresponding to the characteristics of the four management systems. For example, Question 1 asks, "How much confidence is shown in subordinates by superiors?" The four alternatives spread out on the 20-point scale are very little, some, substantial, and very much. Responses more toward the right on the scale receive a higher numerical value and indicate use of a management system more like the participative group System 4 seen by the theory as most effective. According to Likert (1967), the fact that all the scales are in the same direction does not influence the results.

The items in Form S are grouped into the six theoretical dimensions of organizational processes mentioned in the introductory paragraph. Items 1–4 measure leadership; 5–7, motivation; 8–12, communication; 13–15, decision making; 16–17, goal setting; and 18–20, control. Both individual item scores (Golembiewski & Carrigan, 1970a, 1970b; Kavčič et al., 1971) and index scores (Blumberg & Wiener, 1971; Marrow et al., 1967) on the basic dimensions have been used in previous research.

Procedure

As part of a larger study (Butterfield & Farris, 1971) the Likert Organizational Profile was administered in group sessions during visits to Brazilian development finance institutions. On the first administration, instructions were to answer each question twice, once to describe the organization as it was actually and a second time as it should be ideally. On the second administration, 6 to 18 months later, instructions were to describe the organization as it was actually and as it had been previously (at the time of the first administration, although the first administration was not referred to specifically).

Complete data are available from 13 organizations. One was privately owned and the others were either development banks or development finance companies owned by state or regional governments. For 2 banks, there was a 6-month time lag between administrations; for 8 banks, there was a 12-month time lag; and for 3 banks, there was an 18-month time lag after the original administration of the Likert Organizational Profile. These differences in time were designed to look at accuracy of retrospection as a function of the passage of time. It was the authors' impression that the participants in this study were quite interested in the task of completing the questionnaire. The sessions in which the questionnaires were administered seemed to go as well as they usually do in the United States with similar respondents and similar instruments.

Respondents

Respondents were professional employees ranging from nonsupervisors to bank presidents. Sixty-seven per cent were between the ages of 25 and 34; 76% had more than 15 years of schooling; 53% had been with their organizations 4 years or less; and 80% were either lawyers, economists, engineers, or accountants.

Total individual *n*s for the Time 1 and Time 2 administrations were 211 and 256, respectively. They represent a nonrandom sample of about 85% of the professional staff involved with a loan program for small and medium-sized industrial firms. The *n*s within banks varied from 7 to 41, the size of the sample generally reflecting the size of the bank. Size did not influence the results.

Analysis

Except for the factor analyses and hierarchical analyses, data were analyzed using bank mean scores. The bank was selected as the unit of analysis since the Likert Organizational Profile is designed to measure organizational characteristics, and our primary concern is with the management systems of organizations.

RESULTS

Profiles

What management system are these Brazilian development banks actually using, and

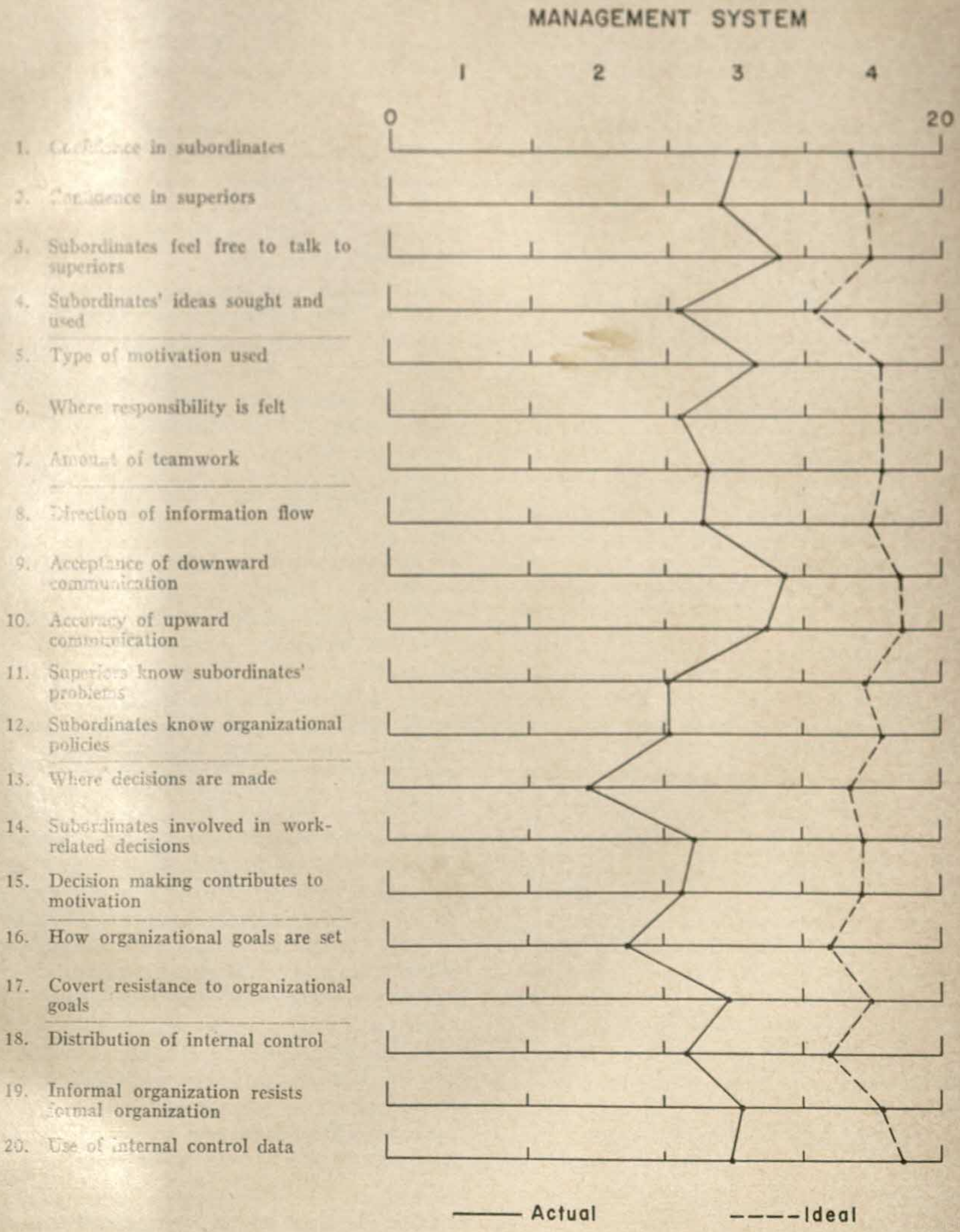


FIGURE 1. Bank mean ideal and actual Likert Organizational Profiles, Time 1 ($n = 13$ banks).

what system would they like to be using ideally? Figure 1 presents the average actual and ideal profiles from the 13 banks, based on the first Likert Organizational Profile administration. It is evident that these 13 financial institutions were actually using a System 3 consultative style of management with System 2 attributes in the area of decision making

TABLE 1

FACTOR ANALYSIS OF IDEAL LIKERT ORGANIZATIONAL PROFILE

Item	Loading
Factor 1: Leadership (5.69,[a] .28[b])	
1) Confidence in subordinates	.80
2) Confidence in superiors	.75
3) Subs feel free to talk to sups	.71
4) Subs ideas sought, used	.63
Factor 2: Resistance (1.85,[a] .38[b])	
17) Covert resistance org. goals	.82
19) Informal org. resists formal	.76
Factor 3: Guidance (1.49,[a] .45[b])	
20) Use of internal control data	.74
9) Acceptance communication down	.71
Factor 4: Informed decision making (1.23,[a] .51[b])	
14) Subs involved decisions related to work	.67
11) Sups know subs' problems	.66
13) Where decisions made	.63
15) D-M contribute motivation[c]	.58
12) Subs know org'l policies[c]	.50
Factor 5: Dispersion of goal setting and control (1.04,[a] .56[b])	
18) Distribution of internal control	.66
16) How org'l goals set	.62
Factor 6: Motivation and communication (.98,[a] .61[b])	
8) Direction of information flow	.76
6) Where responsibility felt	.69
10) Accuracy of communication up	.61

[a] Eigenvalue.
[b] Cumulative proportion of total variance.
[c] Item loads on more than one Factor over .40.

and goal setting. The ideal style was seen as System 4. Analyses of variance indicated essentially no differences among banks on their ideal profiles and significant differences on their actual profiles. Similarly, analysis of variance among three hierarchical levels in the banks (nonsupervisory professionals, supervisors, and top management) showed no difference on ideal profiles, but frequent differences on the actual profiles. Higher organizational levels tended to see their organization as being more toward System 4 than did lower levels. Thus, there was substantial agreement among banks and organizational levels that System 4—participative group—would be the ideal management style, yet the actual system was seen as being more towards consultative or autocratic.

The profiles in Figure 1 are similar to those found in Likert's studies. They suggest that, in terms of Likert's theory, Brazilian development banks may be using management systems not very different from those used by organizations in the United States or even Yugoslavia or Japan.

Factor Analysis of the Likert Organizational Profile

The short form of the Likert Organizational Profile was designed to measure six basic organizational processes: leadership, motivation, communication, decision making, goal setting, and control. Does it, in fact, do this? Factor analyses of the Likert Organizational Profile were carried out to examine its basic structure and to see if there was empirical support for the proposed six dimensions.

Table 1 presents results of the factor analysis of the ideal responses from individual respondents at all organizational levels. A principal-components analysis was used, with varimax rotation for six factors.[3] The first factor is clearly leadership, characterized by mutual confidence and free communication between superiors and subordinates and by utilization of subordinates' ideas. We have given tentative labels to the remaining factors, which are at best mixtures of the dimensions proposed by the theory. Resistance incorporates covert resistance to organizational goals and the presence of an informal organization resisting the formal one. Guidance refers to the open acceptance of communication from above and the use of internal control data for problem solving rather than punishment. Informed decision making unites delegation and involvement in decision making with knowledge of subordinates' problems and of the organization's policies. Dispersion of goal setting and control measures whether the goal setting and control functions are located at the top or spaced throughout the organization. Motivation and communication combines accurate upward communication and horizontal as well as vertical information flow with widespread felt responsibility for achieving the organization's goals.

---

[3] The specific program used was Number BMD03M from the University of California, Los Angeles, Health Sciences Computing Facility (See Dixon, 1967).

In additional analyses it became more difficult to label factors. Note that the analysis of the ideal Likert Organizational Profile responses shown in Table 1 should have stopped with five factors since the eigenvalue for the sixth factor is less than 1.00. A careful examination of Table 1 also indicates that neither Item 5 (type of motivation used) nor Item 7 (amount of teamwork) find a place in any of the six factors.

Additional factor analyses were carried out for the actual responses from the two time periods. However, since results were much less interpretable, they will not be presented here. Controls for organizational level tended to complicate rather than clarify results. The most consistent factors were leadership and resistance, but they emerged in a different order and with a different makeup in each analysis. Usually there were only three or four factors with eigenvalues greater than 1.00. Items 7 and 5 frequently did not appear.

The results in Table 1 represent in a sense the best of what the factor analyses produced, that is, somewhat interpretable factors with at least some relationship to the theory. Repeated analyses produced only fair factor consistency over time, fair overlap between ideal and actual factors, and fair-to-poor comparability across organizational levels. This lack of consistency in factors across different analyses was the characteristic outcome, not the emergence of a large and possibly halo-induced general factor. Other studies (Likert, 1967; Marrow et al., 1967) have alluded to such a halo factor, but none was found in the present case. Since factors were not clear cut nor consistent across the different analyses, the creation of factor scores did not seem justified. Subsequent analyses were based on individual items as well as on a total Likert Organizational Profile score (overall average across the 20 items). The latter seemed justified in view of high intercorrelations among the twenty Likert Organizational Profile items and of Likert's (1967) assertion that the Likert Organizational Profile measures relatively consistent management systems.

Longitudinal Analysis of the Likert Organizational Profile[4]

How reliable are Likert Organizational Profile scores over time? In particular, how accurate are Likert Organizational Profile scores when used retrospectively? The accuracy issue should be of general interest since investigators frequently ask respondents on questionnaires to remember how things were at some point in the past, without knowing in fact how they really had been. In the present study we knew.

Accuracy of retrospection. We might expect that accuracy of retrospection decreases as time increases due to forgetting. Thus, in the present study, the banks in which the Time 2 administration followed the Time 1 administration by 6 months should be the most accurate; retrospective scores from the 18-month time-lag banks should be the least accurate; and the 12-month group of banks should fall in between. However, if retrospection were perfect, the actual and retrospective profiles within each time grouping should be the same. Preliminary visual inspection showed remarkable overlap between the actual (first administration) and retrospective (second administration) profiles within each of the three time groups. The t tests, which were done on an item-by-item basis separately for each time group, confirmed that there were few significant differences between actual and retrospective item scores for any time group. These results are summarized, using mean total Likert Organizational Profile scores, in Table 2. An F test on the average difference score for the

---

[4] All longitudinal analyses were carried out separately for each organizational level. Results were essentially identical. Data presented in the longitudinal section are from respondents at all organizational levels combined.

TABLE 2

Summary: Effect of Time on Retrospection

Time group in months	Mean total Likert Organizational Profile		t	df	p
	Time 1 actual	Retrospective			
6	9.91	10.48	.82	1	*ns*
12	11.54	11.46	−.23	7	*ns*
18	12.32	12.49	.18	2	*ns*

three groupings confirmed that there was no difference in accuracy among the three retrospective periods ($F = 0.27$, $df = 2/10$, $p = ns$).

Test–retest reliability. To look at "real change" as a function of time, similar longitudinal analyses were carried out comparing Time 1 actual Likert Organizational Profile scores with Time 2 actual scores. Results (not shown) were the same as for retrospective scores: time had no effect. Since time differences produced no differences in either retrospective accuracy or change, simple rank-order correlations were calculated using all 13 bank means on the total Likert Organizational Profile scores. Time 1 actual scores correlated $\rho = .63$ with Time 2 retrospective scores and $\rho = .52$ with Time 2 actual scores. With an n of 13, both correlations are significant beyond the .05 probability level using a one-tailed test. The first ρ (.63) might be considered the retrospective reliability of the Likert Organizational Profile. The second ρ (.52) is the test–retest reliability. Though statistically significant, neither is especially high. Of course, the moderate test–retest reliability over the 6–18-month periods is probably due to actual changes in management systems as well as to characteristics of the measuring instrument.

Likert Organizational Profile Scores and Effectiveness and Satisfaction

Likert's theory predicts positive relationships between Likert Organizational Profile scores and such dependent variables as organizational effectiveness and employee satisfaction. That is, the more an organization uses a participative group management system, the better off it will be. To what extent is the theory supported in development banking organizations in Brazil?

TABLE 3

Correlations between Bank Total Likert Organizational Profile Scores and Effectiveness and Satisfaction

Dependent variable	Likert Organizational Profile
Effectiveness	
1. Number of development loans, 1970	−.01
2. Ratio of development loans to total assets, 1969	.07
3. Rankings of overall effectiveness by National Bank experts, 1970	.28
4. Bank self-rating of effectiveness	.78*
Satisfaction	
5. Satisfaction with immediate supervisor	.24
6. Employee satisfaction with bank	.77*
7. Evaluation of bank as place to work	.83*

Note. Rank-order correlations using Time 2 actual Likert Organizational Profile scores and questionnaire data from nonsupervisory professionals only. *N*s range 11–15, including 2 banks not part of longitudinal analysis.

* $p < .01$, one-tailed test.

Table 3 presents correlations between Time 2 actual bank Likert Organizational Profile scores and effectiveness and satisfaction.[5] The first three effectiveness variables are independent "objective" measures of bank performance. The number of development loans, 1970, is a measure of the development productivity of the banks. It represents the number of small and medium industrial loan projects approved for disbursement in 1970. The ratio of development loans to total assets, 1969,[6] indicates the relative emphasis the banks place on development. The greater the proportion of total assets taken up by development loans, the more the bank is fulfilling its primary mission. Rankings of overall effectiveness by National Bank experts, 1970, represents the combined ranks of four judges who independently rank ordered the overall performance of the banks they knew best. (Intercorrelations among the judges' rankings were all statistically significant, ranging from $\rho = .52$ to $.87$). In general, the objective measures were reliable (year-to-year correlations in the .70s and .80s) but only moderately correlated with one another.

---

[5] Time 2 data were selected to maximize N. Missing data and the fact that Time 1 Likert Organizational Profile data were collected at different times for different banks prevented the use of path analysis or of other procedures which might have better indicated the direction of causality.

[6] This ratio includes development loans of all types, not just small and medium industrial projects. The rank ordering of banks on financial data was so highly correlated from year to year in prior years that data on this ratio were not obtained for 1970.

The last four measures in Table 3 are items taken from another questionnaire administered in the banks at the time of the second Likert Organizational Profile administration. They are the employees' own perceptions of how effective their bank is, how satisfied they are with their own supervisor, the extent to which the bank's employees are satisfied rather than dissatisfied, and their evaluation of their bank as a place to work compared with other organizations they know. Since organizational level had a significant effect on these items (higher levels reporting more favorable responses), the results in Table 3 are based on Likert Organizational Profile and questionnaire data from nonsupervisory professionals only.

Table 3 shows that bank scores on the Likert Organizational Profile were not related to objective measures of bank effectiveness, but that they were related in the predicted manner to three of the four questionnaire measures of effectiveness and satisfaction. Contrary to theory, System 4 methods of management were not associated with better organizational performance, but they were associated with employees' feelings about performance and with employee satisfaction. The theory is thus only partially supported.

Note the nature of the one satisfaction item in Table 3 not related to the Likert Organizational Profile. Apparently, employees who described their organization as more towards System 4 were not more satisfied with their own supervisor (nor were they more dissatisfied). One's *immediate* supervisor may not be more satisfying in a System 4 organization, yet employees do feel that the entire organization is more effective, is a better place to work, and has more satisfied employees. It appears that the Likert Organizational Profile can distinguish between more general organization-wide and more immediate work-group characteristics.

DISCUSSION

Development banks are probably among the more modern of Brazilian organizations. They are relatively new and hire a high percentage of young and well-educated employees. Under these conditions, the Likert Organizational Profile appears to be a useful instrument for organization studies in Brazil. Questionnaire administration went quite smoothly, and profiles were remarkably similar to those obtained in the United States. Common stereotypes of Latin American culture would predict that in the more traditional Brazilian organizations we might encounter more difficulties in the administration of the Likert Organizational Profile, and that we might find both actual and ideal profiles toward the authoritative end of the scale. These stereotypes merit empirical testing.

The factor structure of the Likert Organizational Profile does not correspond well to the six hypothesized organizational processes. This is not a serious drawback, but it does suggest that index scores on the subscales should be used with great caution. Likert himself recognizes the difficulty in obtaining factors corresponding exactly to the proposed dimensions (Likert, 1967). He argues that either ideal scores or actual scores from within a single organization will yield cleaner factors. In the present study, we did not have sufficient data to justify factor analysis from within a single organization. Factor scores aside, much can still be done with individual items and especially with a total Likert Organizational Profile score. Using ideal and actual profiles in a feedback program for organizational development has been useful in both Brazil (Butterfield, 1972) and the United States (Blumberg & Wiener, 1971).

The reliability of the Likert Organizational Profile over an extended period of time is a complex question. On the one hand, it seems quite clear that organizational members do a

6·3·81

fairly accurate job in remembering the state of their organization 6, 12, or 18 months into the past. Furthermore, these differences in time have no effect on the accuracy of retrospection. Using the Likert Organizational Profile retrospectively thus appears to be quite legitimate, at least up to a period of 18 months. On the other hand, total scores on the Likert Organizational Profile are only moderately stable over time. The test–retest reliability (.52) obtained in the present study, though statistically significant, is certainly not high by traditional testing standards. However, given the small number of organizations (13) and the rather dynamic state of the development finance field in Brazil, perhaps we ought not to have expected the test–retest reliability to have been much higher.

Results of the correlations between the Likert Organizational Profile and dependent variables are of particular theoretical interest. As predicted, Brazilian development bankers were more satisfied with participative group management systems (although such satisfaction was not dependent on satisfaction with the immediate supervisor) and said such systems would be the ideal. Contrary to theory, organizations tending to use such systems were not more effective in terms of objective performance.

System 4 theory would explain these different results by invoking a time-lag argument. The benefits of using System 4 methods are already apparent in the greater satisfaction of organization members. Such satisfaction usually precedes changes in organizational end-product variables. Improved performance is on the way; it is only a matter of time. Unfortunately, although we do have longitudinal data in the present study, we do not have sufficient data over a sufficient time period to test adequately the time-lag hypothesis.

An alternative explanation is the "manpower competence–performance feedback loop hypothesis" (Farris, 1969; Farris & Butterfield, 1971). This theory holds that the effectiveness of participative methods depends on having competent, experienced employees. When such a manpower base does not exist, as is the likely case in developing countries, participative methods are not likely to be more effective. Greater relative competence resides at the supervisory level, so that closer methods of supervision may be more appropriate as well as more accepted (Farris & Butterfield, 1971). In the present study, for example, satisfaction with immediate supervisor was not significantly related to being in a more participative organization.

These are alternative rather than competing hypotheses. Results of the present study suggest that the Likert Organizational Profile may be a useful instrument for exploring them further, even if System 4 theory is only partially supported in Brazil.

REFERENCES

BLUMBERG, A., & WIENER, W. One from two: Facilitating an organizational merger. *Journal of Applied Behavioral Science,* 1971, 7, 87–102.

BUTTERFIELD, D. A. Pesquisa de ação em desenvolvimento organizacional na América Latina pelo método de levantamento de feedback. *Revista de Administração Pública,* 1972, **6**, 89–99.

BUTTERFIELD, D. A., & FARRIS, G. F. *Organizational development for economic development in Latin America: Results of a four-year action-research program.* (Working Paper No. 543-71) Cambridge: Massachusetts Institute of Technology, Alfred P. Sloan School of Management, 1971.

DIXON, W. F. *Biomedical computer programs.* Berkeley & Los Angeles: University of California Press, 1967.

FARRIS, G. F. Organizational factors and individual performance: A longitudinal study. *Journal of Applied Psychology,* 1969, **53**, 87–92.

FARRIS, G. F., & BUTTERFIELD, D. A. Are current theories of leadership culture-bound? An empirical test in Brazil. In, Contemporary Developments in the Study of Leadership. Symposium presented at Southern Illinois University, Carbondale, April 1971.

FOUNDATION FOR RESEARCH ON HUMAN BEHAVIOR. *Sample kit: Profile of organizational characteristics.* Ann Arbor, Mich.: Institute for Social Research, 1967.

GOLEMBIEWSKI, R. T., & CARRIGAN, S. B. The persistence of laboratory-induced changes in organization styles. *Administrative Science Quarterly,* 1970, **15**, 330–340. (a)

GOLEMBIEWSKI, R. T., & CARRIGAN, S. B. Planned change in organization style based on the laboratory approach. *Administrative Science Quarterly,* 1970, **15**, 79–93. (b)

KAVČIČ, B., RUS, V., & TANNENBAUM, A. S. Control, participation, and effectiveness in four Yugoslav industrial organizations. *Administrative Science Quarterly*, 1971, **16**, 74–86.

Likert, R. *New patterns of management*. New York: McGraw-Hill, 1961.

Likert, R. *The human organization*. New York: McGraw-Hill, 1967.

Likert, R. The relationship between management behavior and social structure—improving human performance: Better theory, more accurate accounting. Symposium (C-3) presented at the meeting of the Conseil International Pour L'Organisation Scientifique, Tokyo, 1969.

Marrow, A. J., Bowers, D. G., & Seashore, S. E. *Management by participation*. New York: Harper, 1967.

(Received July 24, 1972)

Manuscripts Accepted for Publication in the

Journal of Applied Psychology

Replication of White-Collar-Blue-Collar Differences in Sources of Satisfaction. Edwin A. Locke (College of Business and Public Administration, University of Maryland, College Park, Maryland 20742) and Thomas C. Harris.

Relationships between Performance and Satisfaction under Contingent and Noncontingent Reward Systems. Gerald A. Kesselman (Department of Psychology, Ohio State University, 404C West 17th Avenue, Columbus, Ohio 43210), Michael T. Wood, and Eileen L. Hagen.

Identifying Responsive Inner City Policemen. Rita Mae Kelly (10 Wilson Terrace, Elizabeth, New Jersey 07208) and Martin G. Farber.

Factors Influencing Disciplinary Judgments. Benson Rosen (Graduate School of Business Administration, University of North Carolina, Chapel Hill, North Carolina 27514) and Thomas H. Jerdee.

Measuring Organizational Communication. Karlene H. Roberts (School of Business Administration, University of California, Berkeley, 350 Barrows Hall, Berkeley, California 94720) and Charles A. O'Reilly III.

Factor Analysis of Air Passenger Reactions to Skyjacking and Airport Security Measures as Related to Personal Characteristics and Alternatives to Flying. Robert J. Wherry, Sr. (Department of Psychology, Ohio State University, 404C West 17th Avenue, Columbus, Ohio 43210), Michael R. Cooper, and Charles E. Boltwood.

Repetition of Highly Similar Messages and Attitude Change. J. Lee McCullough (Hay Associates, 1845 Walnut Street, Philadelphia, Pennsylvania 19103) and Thomas M. Ostrom.

Influence of Technical Knowledge on Position Error in Ranking. Edwin E. Wagner (Department of Psychology, University of Akron, Akron, Ohio 44325) and Thomas O. Hoover.

Development of a Biographical Inventory Blank to Predict Faking on Personality Tests. Joel Lefkowitz (Baruch College, City University of New York, 17 Lexington Avenue, New York, New York 10010) and Jeffrey Cohen.

Components of Perceived Risk: A Cross-Validation. Jacob Jacoby (Department of Psychological Sciences, Purdue University, West Lafayette, Indiana 47907), Leon B. Kaplan, and George J. Szybillo.

Cluster Analyses of Interest Inventory Scales as Tests of Holland's Occupational Classification. Dean H. Nafziger (Center for Study of Social Organization of Schools, The Johns Hopkins University, Baltimore, Maryland 21218) and Samuel T. Helms.

(*Continued on page 30*)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 24–30

EFFECTS OF A MERGER OF A SMALL AND A LARGE ORGANIZATION ON MEMBERS' BEHAVIORS AND EXPERIENCES [1]

ALLAN W. WICKER [2] AND CLAUDIA E. KAUMA
Claremont Graduate School *University of Wisconsin–Milwaukee*

After the merger of two local churches consisting of 419 and 961 members, 118 members formerly from the smaller church and 115 members formerly from the larger church reported their support for activities in both the prior and merged churches, and their feelings about participating in the merged church. Church support prior to the merger did not differ for the two groups. Results indicated that both groups declined in support after the merger, but the decline was greater for members from the small church. The small church sample felt less close to the merged church, less obligated and less welcome to participate in its activities.

A number of recent studies have suggested that participation in voluntary activities of an organization is related to the organization's size: members of small schools and churches have been found to assume more positions of responsibility, to enter more different kinds of activities, and in general to show higher levels of support for the activities than members of large schools and churches (Baird, 1969; Barker & Gump, 1964; Wicker, 1969a, 1969b; Wicker & Mehler, 1971). These behavioral differences have been accompanied by differences in subjective experiences and social norms; members of the small organizations have reported greater feelings of responsibility, involvement, and obligation to participate, and they have shown greater approval of high levels of participation (Barker & Gump, 1964; Wicker, 1968, 1969b; Wicker & Mehler, 1971).

The research cited necessarily employed correlational designs in which investigators selected organizations on the basis of membership size, and then obtained measures of participation and experiences from a sample of members in each organization. This procedure, particularly for the studies of churches, leaves open the question of direction of effect. It seems plausible that church members might actively select a church of a size that makes demands compatible with their willingness to expend effort. The present investigation deals with this problem by taking advantage of a natural experiment—the merger of two churches. Thus it was possible to compare the behavior and experiences of the same individuals when they were members of organizations differing in size.

The two merging churches were located within six blocks of one another in a working-class area of Milwaukee, Wisconsin. The denomination of the smaller church was Evangelical United Brethren; it had 419 members. The larger church was Methodist, and had 961 members. These two local congregations voted to merge after the national organizations of these denominations merged to form the United Methodist Church. The building occupied by the smaller church was abandoned after the merger. Both pastors of the former churches shared in conducting worship services in the merged church. Joint committees were organized to integrate the members into the various activities of the merged church. In the first year, the officers of the various organizations in the merged church included 98 from the small church and 200 from the large church.

---

[1] This report is based on a master's degree thesis submitted to the Department of Psychology, University of Wisconsin–Milwaukee, by the second author. The first author supervised the research. The authors are indebted to Alan Baron, Anthony Fazio, and J. Richard Hackman for helpful comments at various stages of the present research, and to Martin Thomas and Galen Luebke and members of their congregation for cooperation in the study.

[2] Requests for reprints should be sent to Allan W. Wicker, Graduate Faculty in Psychology, Benezet Psychology Building, Claremont Graduate School, Claremont, California 91711.

Unfortunately, the selectivity problem was not wholly eliminated in the present study. Seventy-one of the members from the smaller church did not continue membership in the merged church. Of these, approximately one fourth stated they did not wish to affiliate with the Methodist denomination, one fourth had been inactive members, and one half lived in outlying areas and saw the transition as an opportunity to join churches closer to their homes. This attrition of members from the smaller church was not judged a serious problem, however, since it was clearly not the case that the most active members of the smaller church were discontinuing membership.

A number of specific expectations for the present study were derived from the assumption that organization size determines, in a causal sense, members' levels of participation in organization activities and their feelings of identification with the organization, its activities, and its members. These expectations were as follows: (*a*) Prior to the merger, members of the small church would have shown greater levels of support for their church than members of the large church, while after the merger, the two groups should not differ in their levels of support, nor in their feelings regarding support (e.g., feeling obligated, welcome to participate). (*b*) The degree to which subjects were acquainted with members from the other church would be the same for both groups. (*c*) While members from both groups would show a decline in church support as a result of the merger, this decline should be greater for the members of the small church since they experienced a greater increase in church size. (*d*) The latter group would report, after the merger, a greater decline in feelings of pressure to participate and in feelings of closeness to the church than would members from the large church.

METHOD

Subjects

Names of subjects contacted were drawn from the membership list of the merged church. All married couples formerly from the small church ($n = 65$) and an equal number of randomly selected couples formerly from the large church were sent a letter from the church pastors requesting their participation in the study. The letter introduced the investigators, stressed the importance of the study to the church, and stated that members would be contacted for a telephone interview within the next few weeks.

An attempt was made to interview both members of every couple. Repeated attempts were made to contact subjects not initially available. Of 130 former members (i.e., 65 couples) of the small church, 118 (91%) eventually provided data and 12 (9%) refused to participate. Of 130 former members (also 65 couples) of the large church, 115 (88%) provided data, 2 (2%) could not be reached, and 13 (10%) refused to participate. The sex distribution of the participating subjects from each church did not differ (50% males and females in each sample).

Telephone Interview

Telephone interviews, conducted 11 months after the merger, dealt with members' participation in and feelings about church activities. (Complete copies of the interview schedules are provided in Kauma, 1972.) The interviewer's introductory statement assured subjects that their responses would be confidential, and that although a report of the study would be made to church officials, it would contain only group data.

It had been determined that there were 13 activities common to the small, large, and merged churches; these included religious study groups, administrative board meetings, church picnics, church school classes, choir rehearsals, and several other organized activities. Subjects were asked if they had attended each activity either during the year before the merger or during the year following the merger. If a subject had attended an activity at either time, he was asked if he had had any job or responsibility, such as serving as an officer or helping out in any other way; on the basis of this information, his level of participation was categorized as (*a*) member (attending but having no particular responsibility), (*b*) worker (helping but not having a major responsibility), or (*c*) leader (taking a position which involved some control and authority in the activity).

Subjects were also asked to indicate, for both before and after the merger: (*a*) the number of hours per month they spent in all church activities, excluding Sunday worship service, and (*b*) the number of times per month they attended Sunday morning worship service. Other questions were (*a*) whether subjects felt obligated to participate in each of six activities of the merged church (including choir, men's or women's organizations, couples club, and others), (*b*) whether they felt welcome to participate in the same six activities, (*c*) whether they felt under more, less, or about the same amount of pressure to participate in church activities as a result of the merger, (*d*) whether they felt closer to the merged church, not so close, or about the same as they felt toward the prior (small or large) church, and (*e*) whether they were acquainted with five persons named by the interviewer. The persons named were former officers of the small or large church; members from the small church were given the names of the officers from the large church and vice versa.

Data from Church Records

Data on amount of money contributed were available for couples in the small, large, and merged churches. At the beginning of each fiscal year, members are asked to pledge the amount of money they are willing to give during the year. Contributions are placed in envelopes with an identifying number, and records are kept for each donor unit (family). Pledge data were available for the merged church only. Worship service attendance data on individuals were available only for the merged church. These data were based on signup sheets which were passed along the pews at each service. Data on the number of persons attending Sunday worship service per Sunday (based on head counts) were available for the small, large, and merged churches. This last measure was thus based on all members of the churches, and not merely the present sample.

RESULTS

Support for Church Activities

A 2 (church condition, prior or merged) × 2 (church size, large or small) repeated measures analysis of variance was applied to each of the five support measures in Table 1. It was expected that, for each measure, there would be a significant main effect due to church condition (lower support in the merged church than in the prior churches) and a significant interaction (a greater difference between prior and merged conditions for members from the small church than for members from the large church). The main effect of size of church from which members came was not expected to be significant, since it represented the mean across the two church conditions, prior and merged.

The data in Table 1 show that the present predictions were partially confirmed: for four of the five measures, the level of support was significantly greater in the prior church than in the merged church, and for two of the five measures, there were significant interactions. The plots of these interactions (Figure 1) show the expected greater decline in support by members from the small church. None of the measures showed significant main effects due to the size of the church from which subjects came.

In another type of analysis of the support data, composite indices of support were derived and *t* tests were applied to three comparisons of interest: (*a*) in the prior church condition, support by members of the small church versus support by members of the large church (expectation: small > large), (*b*) in the merged church condition, support by members from the small church versus support by members from the large church (expectation: small = large), and (*c*) the difference between support in the prior and merged churches by members from the small church versus the same difference for members from the large church (expectation: small > large).

Before the composite scores were derived, the four support measures available on individual subjects in both church conditions (i.e., the first four measures in Table 1) were intercorrelated. Data for members from the small and large churches were pooled and separate matrices were computed for the prior church and the merged church. For each church condition, all coefficients were positive and relatively high: (*a*) In the prior church condition, the range was from .40 to .94, with a median of .67. (*b*) In the merged church condition, the range was from .39 to .88, with a median of .63. It was therefore judged legitimate to combine the individual measures into an overall index of support.

Composite indices of support were generated in the following manner: for each of the four measures, subjects' scores were converted into standard scores; the four standard scores were then summed to yield a support composite for each subject. Three different composite scores were calculated for each subject, one for each comparison of interest: The first was based on the distributions of support measures in the prior churches, the second on the distribution of support measures in the merged church, and the third on the distributions of difference scores obtained by subtracting each support measure in the merged church from the corresponding measure in the prior church.

The expectation that for the prior church condition, members of the small church would show higher levels of support for church

TABLE 1

MEANS AND *F* RATIOS FOR MEASURES OF CHURCH SUPPORT

Measure of support	Prior church		Merged church		*F* ratios		Interaction
	Small	Large	Small	Large	Small vs. large	Prior vs. merged	
No. activities attended	2.2	1.9	.9	1.3	<1	112.31*	11.32*
No. activities in which subjects were workers or leaders	1.8	1.4	.5	.8	<1	99.40*	14.97*
No. hours per month spent in activities	5.8	6.6	2.7	4.7	1.67	23.47*	1.15
No. times per month subjects attended Sunday worship service	2.9	2.7	2.3	2.3	<1	30.39*	<1
Amount of money contributed per couple	$151	$138	$140	$144	<1	<1	1.89

Note. For all analyses $n = 118$ for small and $n = 115$ for large, except for amount of money contributed per couple for which $n = 60$ for small and $n = 55$ for large.

* $p < .01$.

activities than members of the large church was not supported (small church composite index $\bar{X} = .16$, large church $\bar{X} = -.17$; $t = .74$, $df = 231$, *ns*). Also, the expectation that the two groups of subjects would not differ in their levels of support in the merged church was disconfirmed at a borderline level of significance; members from the small church tended to show less support ($\bar{X} = -.38$) than members from the large church ($\bar{X} = .39$; $t = 1.77$, $df = 231$, $p < .10$). As expected, the decline in support after the merger was significantly greater for members from the small church ($\bar{X} = .58$) than for members from the large church ($\bar{X} = -.60$; $t = 3.15$, $df = 231$, $p < .01$).

An additional measure of support in the small, large, and merged churches was worship service attendance based on head counts taken at the services. Across the four quarters which immediately preceded and followed the merger, the percentage attendance[3] (Wicker, McGrath, & Armstrong, 1972) was greater in the small church than in the large church (38% versus 29%; $t = 8.60$, $df = 3$, $p < .01$) or in the merged church (38% versus 30%; $t = 16.20$, $df = 3$, $p < .01$). However, the percentages of members who attended services in the large church and in the merged church were nearly identical (29% versus 30%).

Two measures of church support were available only for the merged church: (*a*) amount of money pledged to the church, and (*b*) frequency of attendance at Sunday worship service based on the sign-up sheets. As expected, members from the small and large churches did not differ significantly on either measure.

Feelings about Supporting Church Activities

Responses to the question regarding feelings of closeness toward the merged church as compared to the prior church showed the expected pattern. A greater percentage of members from the small church reported feeling less close than did members from the large church (50% versus 33%), while more members from the large church reported no difference in feelings of closeness (57% versus 45%) and feelings of increased closeness (10% versus 5%). The chi-square for the 3 × 2 contingency table = 7.89, $df = 2$, $p < .05$.

Contrary to expectation, members from the small and large churches did not differ markedly in responses to the question asking

---

[3] Some problems in using the statistic, number of persons attending/number of persons on membership rolls, as an index of members' support have been noted by Wicker, McGrath, and Armstrong (1972): It is relatively unstable for organizations having small memberships, and visitors are generally counted along with members in the attendance tally. Although the instability problem is probably not severe for churches as large as those in the present study, the data on percentage of attendance should be interpreted and evaluated in the context of the other support measures reported.

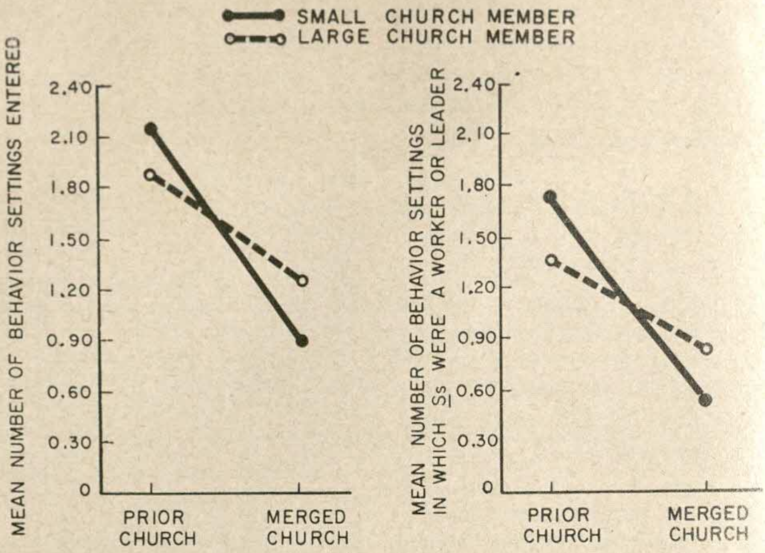


Figure 1. Members' support for church activities: Interactions between size of church from which members came and prior-merged church conditions.

them to compare the pressure to participate in church activities in the merged church with the pressure in the prior church. A greater percentage of the small church members reported less pressure (21% versus 15%) and more pressure (7% versus 2%), while more members of the large church reported the same amount of pressure (83% versus 73%). The chi-square for the 3 × 2 contingency table = 5.83, $df = 2$, $p < .10$).

Although no differences were expected between persons from the small and large churches in feelings of obligation or feelings of being welcome to participate in activities of the merged church, the two groups did differ. Members from the small church felt olbigated to participate in fewer of the six named activities of the merged church than did members from the large church ($\bar{X}$s were .6 and 1.1, respectively; $t = -2.87$, $df = 231$, $p < .01$). The same pattern was obtained for feelings of being welcome to participate in the same six activities: small church $\bar{X} = 4.1$; large church $\bar{X} = 4.8$, ($t = -3.80$, $df = 231$, $p < .01$). Moreover, significantly more members from the small church said they did not feel obligated to participate in any of the activities (69% versus 49%, $\chi^2 = 10.24$, $df = 1$, $p < .001$). A large percentage of members from the small church also reported that they did not feel welcome to participate in any of the settings (13% versus 3%, $\chi^2 = 8.67$, $df = 1$, $p < .01$).

Acquaintanceship with Church Members

Members from the two churches also differed in their acquaintance with five former officers from the other church. The mean number of former officers of the large church known by members of the small church (1.9) was greater than the number of former officers of the small church known by members of the large church (1.3; $t = 3.27$, $df = 231$, $p < .01$). Also, a smaller percentage of members from the small church reported that they did not know any of the former officers (28% versus 45%, $\chi^2 = 7.41$, $df = 1$, $p < .01$).

Discussion

The pattern of the present results, while suggesting that organization size does influence members' participation and experiences, also suggests that the additional factor of site chosen for the merged church may have played an important role.

The strongest findings supporting the hypothesis that increases in organization size produce reduced levels of participation and lesser feelings of identification with the organization were the following: (*a*) Support for the (smaller) prior churches was greater than for the (larger) merged church. (*b*) Decrements in support following the merger generally were greater for the members from the small church than for members from the large church. (*c*) Feelings of being less close to the merged church than to the prior church were more prevalent for members from the small church. However, members from the small church did not report a greater decline in pressure to participate as a result of the merger than did members from the large church, nor did the former group show generally higher levels of church support in the year preceding the merger.

A number of differences obtained between the two groups as members of the merged church suggest that organization size, by itself, is an inadequate explanation for all of the present results. Although the size of the merged church was obviously the same for both groups, members from the small church tended to show lower levels of support, and felt welcome and obligated to participate in fewer activities than members from the large church. The apparent weaker identification with the merged church by members from the small church may be due to the fact that for them, the merger involved leaving a familiar building and migrating to a different site (that of the large church) which was already occupied and functioning. And although there were attempts to integrate the members from the small church into the merged church by holding joint social and organizational meetings and by electing a number of small church members to official positions in the merged church, many of them may have regarded themselves (and have been regarded) as new members of an already functioning church.

Two kinds of additional evidence are consistent with the above interpretation: (*a*) During the telephone interviews, a number of members from the small church spontaneously expressed feelings of alienation from the merged church and some members from the large church indicated they felt their group had not done enough to draw in the members from the small church. (*b*) Results from an earlier study (Wicker & Mehler, 1971) that contrasted new and established members in two other churches showed a pattern very similar to that obtained in the small-large comparisons of the present study. Wicker and Mehler found that new members showed little or no difference from established members in participation in the more open kinds of church activities, such as Sunday worship service attendance and church contributions, both of which are available even to visitors. But the new members showed considerably lower rates of participation than established members in the various other church activities, such as business meetings, social gatherings, and so on. These differences precisely parallel the present differences between members from the small and large churches in the merged church (Table 1).

Finally, the following result may at first glance appear to be inconsistent with the above discussion: members from the small church had a greater familiarity with former officers from the large church than did members from the large church with former officers from the small church. If the members from the small church identified less closely with the merged church, then why would they be more familiar with former officers of the large church? Possibly because the large church's former officers were more active and visible in the merged church than were the former officers of the small church.

REFERENCES

BAIRD, L. L. Big school, small school: A critical examination of the hypothesis. *Journal of Educational Psychology,* 1969, **60**, 253–260.

BARKER, R. G., & GUMP, P. V. (Eds.) *Big school, small school: High school size and student behavior.* Palo Alto, Calif.: Stanford University Press, 1964.

KAUMA, C. E. Effects of the merger between a small and large church on members' behaviors and subjective experiences. Unpublished master's thesis, University of Wisconsin–Milwaukee, 1972.

WICKER, A. W. Undermanning, performances, and students' subjective experiences in behavior settings of large and small high schools. *Journal of Personality and Social Psychology,* 1968, **10**, 255–261.

WICKER, A. W. Cognitive complexity, school size, and participation in school behavior settings: A test of the frequency of interaction hypothesis. *Journal of Educational Psychology*, 1969, **60**, 200–203. (a)

WICKER, A. W. Size of church membership and members' support of church behavior settings. *Journal of Personality and Social Psychology*, 1969, **13**, 278–288. (b)

WICKER, A. W., MCGRATH, J. E., & ARMSTRONG, G. E. Organization size and behavior setting capacity as determinants of member participation. *Behavioral Science*, 1972, **17**, 499–513.

WICKER, A. W., & MEHLER, A. Assimilation of new members in a large and a small church. *Journal of Applied Psychology*, 1971, **55**, 151–156.

(Received October 16, 1972)

(*Continued from page 23*)

Brightness Contrast Effects on Recall of Projected Highway Sign-Type Stimulus Material. Robert W. Swezey (11316 Links Court, Reston, Virginia 22090).

Landscape Office Attitudes: Reflections of Perceived Degree of Control over Transactions with the Environment. Michael McCarrey (Personnel Psychology Centre, Public Service Commission, 28th Floor, Tower "C," Place de Ville, Ottawa, K1A OM7, Ontario, Canada), Lyle Peterson, Shirley Edwards, and Paul von Kulmiz.

An Empirical Comparison of the Relative Effects of Rater Response Biases on Three Rating Scale Formats. Robert F. Burnaska (Corporate Personnel Research, General Electric Company, College Boulevard, Portmouth, Virginia 23705) and Thomas D. Hollmann.

Intrinsic, Extrinsic, and Risk Propensity Factors in the Work Motivation of Teachers, Educational Administrators, and Business Managers. Cecil Miskel (School of Education, University of Kansas, Bailey Hall, Lawrence, Kansas 66045).

Relationship of Role Conflict and Role Ambiguity to Job Involvement Measures. W. Clay Hamner (Graduate School of Business Administration, Eppley Center, Michigan State University, East Lansing, Michigan 48824) and Henry L. Tosi.

Type of Transmittal Letter and Questionnaire Color as Two Variables Influencing Response Rates in a Mail Survey. Michael T. Matteson (College of Business Administration, University of Houston, Cullen Boulevard, Houston, Texas 77004).

Predicting Alumni Financial Donations. Warren S. Blumenfeld (School of Business Administration, Georgia State University, 33 Gilmer Street Southeast, Atlanta, Georgia 30303) and Patricia L. Sartain.

Achievement Motivation among Managers of Small Businesses, Corporation Managers, and Business Students. John C. South (School of Business and Administration, Duquesne University, Pittsburgh, Pennsylvania 15219).

Extensions to a Path–Goal Theory of Motivation. Martin G. Evans (Faculty of Management Studies, University of Toronto, 246 Bloor Street, Toronto M5S IV4, Canada).

Journal of Applied Psychology
1974, Vol. 59, No. 1, 31-37

ROLE MOTIVATION THEORY OF MANAGERIAL EFFECTIVENESS IN SIMULATED ORGANIZATIONS OF VARYING DEGREES OF STRUCTURE

JOHN B. MINER[1]
Georgia State University
DOROTHY N. HARLOW
University of South Florida

JOHN R. RIZZO
AND
JAMES W. HILL
Western Michigan University

This study was conducted to test the hypothesis that the role motivation theory of managerial effectiveness is applicable in highly structured organizational contexts of the bureaucratic type, but not applicable in low structure contexts of the professional type. Simulated organizations with student members were used—a facsimile of a large business firm at Western Michigan University and a series of independent project teams at the University of South Florida. The results generally support the hypothesis, with managerial motivation being significantly related to promotion decisions in the high structure situation, but not in the low structure situation.

One approach to the study of human motivation in the organizational context involves an emphasis on the motivational requirements of tasks and of informal and formal role requirements, rather than on the specific needs and motives of individuals, which are the central focus of the various content theories of motivation. Such a role-based approach has as its major concern the sum total of motivational inputs needed to satisfy job requirements. What is important is that these motivational inputs make a difference in the performance of a particular kind of job rather than that they be of a particular type, such as the need for self-actualization or intrinsic job satisfaction.

To date this role-based approach has been developed most fully in the areas of entrepreneurship (McClelland & Winter, 1969) and the management of bureaucracies (Miner, 1965a). However, as has been noted elsewhere (Miner & Dachler, 1973), the approach could be extended to cover a whole network of key jobs in varying organizational contexts. There is thus the possibility of creating a highly parsimonious but comprehensive theory of motivation at work, which would provide a base for meaningful organizational actions, long before the total complexity of human motivation is understood.

A basic problem in the development of such a theory relates to the generality or breadth of job definitions. Research to date indicates that a theory originally developed to apply to managerial positions in the context of a large highly structured, hierarchic organization cannot be extended to less structured professional organizations where face-to-face relationships predominate (Miner, 1971). Thus the degree of structuring in the organizational environment where the job is performed appears to be an important variable in defining job categories for purposes of developing role-based theories of motivation at work. The present research attempts to shed further light on this matter by moving from the field situation typified by ongoing organizations to simulated situations where a greater degree of control of relevant variables is possible.

Role-Motivation Theory of Managerial Effectiveness

The theory which provided a stimulus to the studies reported in this article posits six types of role prescriptions for managerial jobs. These role prescriptions are assumed to be primarily of an informal nature in most organizations, although the fact that they are formally stated would not make the theory less applicable. The role requirements are hypothesized to have a high probability of occurrence across a broad range of managerial jobs and their presence is expected to be positively related to organizational goal attainment and thus to contribute to organizational effectiveness. The theory as originally stated was viewed as applying to "business firms organized in accordance with the scalar principle" (Miner, 1965a).

---

[1] Requests for reprints should be sent to John B. Miner, Department of Management, Georgia State University, 33 Gilmer Street, S.E., Atlanta, Georgia 30303.

The six role requirements and their accompanying motivation may be described as follows:

(*a*) Managers are typically expected to behave in ways which do not provoke negative reactions from their superiors. In order to represent his group upward in the organization and to obtain support for his actions a manager should maintain good relationships with those above him. It follows that he should have a generally positive attitude toward those holding positions of authority.

(*b*) There is, at least insofar as peers are concerned, a strong competitive element built into managerial work. A manager must compete for the available rewards, both for himself and for his group. If he does not, he may lose ground as he and his function are relegated to lower and lower status levels. Certainly without competitive behavior rapid promotion is very improbable. Thus, managers must characteristically strive to win for themselves and their groups. In order to meet this role requirement a person should be favorably disposed toward engaging in competition.

(*c*) There is a marked parallel between the requirements of the managerial role and the more general demands of the masculine role as it is defined in our society. Both a manager and a father are supposed to take charge, to make decisions, to take such disciplinary action as may be necessary, and to protect others. Thus, one of the more common role requirements of the managerial job is that the incumbent behave in an assertive masculine manner. Even when women are appointed to managerial positions they are expected to follow an essentially masculine behavior pattern. It follows that a desire to meet the requirements of masculinity as it is typically defined will generally lead to success in meeting certain role prescriptions of the managerial job as well.

(*d*) This is the requirement that the manager exercise power over subordinates and direct their behavior in a manner consistent with organizational, and presumably his own, objectives. He must tell others what to do when this becomes necessary, and enforce his words through appropriate use of positive and negative sanctions. The individual who finds such behavior difficult and emotionally disturbing would not be expected to meet this particular role prescription.

(*e*) The managerial job requires a person to stand out from his group and assume a position of high visibility. He cannot use the actions of his subordinates as a guide for his own behavior as a manager. Rather he must deviate from the immediate group and d things that will inevitably invite attention discussion, and perhaps criticism from thos reporting to him. When this prospect i viewed as unattractive, when the idea o standing out from the group, of behaving in different manner, and of being highly visibl elicits feelings of unpleasantness, then behavior appropriate to the role will occur muc less often than would otherwise be the case.

(*f*) The final role requirement refers to th process of getting the work out and keepin on top of routine demands. The things tha have to be done must actually be done. Ther are administrative requirements of this kin in all managerial work, although the specifi activities will vary somewhat from one situation to another. To meet these prescriptions manager must at least be responsible in dea ing with this type of routine, and ideally h will gain some satisfaction from it (Mine 1965a).

In order to test this theory the Miner Se tence Completion Scale was developed. provided measures of motivation relevant the six role prescriptions. The Miner sca yields two total scores and seven subsca scores (two in the area of competitiveness These subscales have been designated autho ity figures, competitive games, competiti situations, masculine role, imposing wishe standing out from group, and routine admi istrative functions.

Tests of the theory using the MSCS ha

produced positive results in the types of organizations and under the conditions where it was originally considered applicable (Gantz, Erickson, & Stephenson, 1971; Miner, 1965a). Criteria used have included supervisory performance ratings, peer ratings, promotion rates, and managerial level. However, the theory has not received support when extended to less structured organizations such as a management consulting firm which uses the professional form and three small school distincts where the administrative process is largely of a face-to-face nature (Miner, 1971). Also, the criterion must be one of *managerial* performance; in the research and development context when promotion is heavily weighted with creative and professional components negative correlations may in fact be obtained (Gantz, Erickson, & Stephenson, 1972; Miner, 1965b).

In the hierarchically structured organization, which relies heavily on formalized procedures, relationships and communication of a vertical nature are of major importance; the very nature of the bureaucratic form makes this so. Thus it does not seem surprising that the Miner Sentence Completion Scale subscales dealing with role prescriptions in the areas of authority relationships, competition with peers, and exercising power over subordinates have contributed the most consistently significant correlations with criteria, although all the other subscales have produced significant results on occasion. In small face-to-face and professional organizations, on the other hand, horizontal relationships and communication tend to be emphasized to a much greater extent, relative to the vertical; the Miner Sentence Completion Scale variables should make little difference under such circumstances.

Furthermore, effectively functioning organizations should place high value on and bestow rewards for behavior that is in accordance with the role requirements of their particular organizational forms. Thus, both role performance evaluations and role assignments (managerial level) should be predictable from the variables of role–motivation theory. In addition, to the extent individuals are aware of the requirements of managerial work in hierarchic organizations and possess the motives required to perform in accordance with them, these individuals would be expected to seek out managerial positions in order to satisfy their strong motives. There is evidence that this does in fact occur (Miner & Smith, 1969).

Given these lines of reasoning and the results obtained from research on actual organizations, the role motivation theory of managerial effectiveness should prove applicable to any reasonably large organization of the administrative or bureaucratic type, and the very fact of structuring an organization in this manner should tend to elicit a positive value toward the role prescriptions of the theory. One method of investigating this hypothesis is to test the theory in temporary simulated organizations having varying degrees of vertical structure. Where structuring is greater one would expect the Miner Sentence Completion Scale measures to correlate more highly with criterion variables. The approach involved here in moving from the field to the simulated situation is essentially the same as that used by Ghiselli and Lodahl (1958) in establishing the construct validity of the Ghiselli Self-Description Inventory (Ghiselli, 1971).

METHOD

Low-Structure Situation

The low-structure simulation used student teams that worked on various case projects throughout the spring quarter of 1972 at the University of South Florida. Each project team worked independently of the others on a separate and distinct problem; the teams were drawn from three different management courses. The groups formed on a voluntary basis, rather than with an assigned membership, and in some cases they formed well in advance of the start of the course. The amount of interaction was sizable with some teams spending as much as 40 hours per week working together. At least 50% of the course grade derived from this group effort. There was no organizational structure above the project team level and there was no specification of a division of labor on either a horizontal or vertical basis within the teams. To the extent such internal structuring appeared it was entirely emergent.

In all there were 19 teams studied with memberships ranging from 4 to 7. The total number of individuals in the teams was 89. Of these 72 completed the Miner Sentence Completion Scale during the spring quarter. These 72 constitute the sample for purposes of study.

At the end of the quarter when all projects were completed the team members filled out a brief questionnaire based on their experiences in working quite closely together over a two-month period. The questionnaire was designed to get at the nature and extent of emergent leadership; it was worded as follows:

> Assume you are a manager of the people listed above. You have received a request from your immediate superior to recommend one of them for a promotion as a manager of another group, i.e., he is not your replacement. What would your recommendation be in terms of who you think would be a good leader?

The assumption was that if role motivation theory applies at all, in such low structure situations, it should differentiate between those nominated for leadership positions as a result of their behavior in the project team context and those viewed as lacking leadership potential.

Emergent leadership potential scores were calculated by giving 2 points for a first choice and 1 point for a second choice. To correct for differences in group size and because only 72 of the 89 group members completed the questionnaire, the score for each person was computed as a percentage of the total number of points that could have been obtained. Values of 60 or higher were taken as indicative of managerial potential. Using this criterion 7 groups had a single leader and 12 had dual leadership. Of these, 4 individuals did not complete the Miner Sentence Completion Scale; the total number of emergent leaders was thus 27. Their mean score was 80.

The dual leadership phenomenon appeared to be a function of a certain amount of clique formation within the groups. As would be expected the single leaders had a higher leadership score, 94 versus 76. However, the slight trend toward higher total Miner Sentence Completion Scale scores for the single leaders did not approximate significance and thus combining of the single and dual leaders for purposes of the analysis seemed justified. There was no significant difference between the mean total Miner Sentence Completion Scale scores of members of one-leader and of two-leader groups.

In addition to comparing the Miner Sentence Completion Scale scores of the 27 leaders with those of the 45 nonleaders, a further differentiation was made within the latter group. Those with leadership scores in the 20 to 59 range were designated as possible leaders; the mean score for these 20 individuals was 42. Those with the lowest scores were designated definite nonleaders; these 25 individuals had a mean score of 6. None had more than a single second choice nomination.

High-Structure Situation

The high-structure simulation study used a student-managed simulated organization called the Management Development Corporation operated by individuals enrolled in management courses at Western Michigan University. Management Development Corporation was a six-level divisionalized organization established for training purposes. The hierarchical form was intended to provide peer, superior, and subordinate role experience.

The students worked on actual current problems. Often the output was in the form of services delivered to the management department as client (computer printouts, data on several classes associated with grade records, analyses of department studies, etc.). It could also be in the form of final plans (activities, costs, scheduling of resources and the like toward the accomplishment of some objective to which the plan addressed itself); information to relevant publics about Western Michigan University, the management department, and the Management Development Corporation (brochures, newspaper articles, newsletters, etc.); and personnel and administrative records related to the Management Development Corporation itself. Operating procedures tended to vary somewhat from semester to semester depending on the problems dealt with and the particular management group.

The organization was headed by a president who reported to a three-man faculty board. The board reviewed and approved progress toward objectives and received final oral reports at the end of the semester. Performance appraisals were made by the students themselves without any direct faculty involvement. These appraisals were in the form of some type of superior rating and were subsequently converted into course grades.

Subordinate to the president were varying numbers of vice presidents, general managers, department managers, project managers, and staff analysts depending on course enrollments. The structure was in the form of a pyramid with decreasing numbers at each higher level. Leadership was appointive with promotion decisions having been made by superiors. The organization could and did, promote at any time, but larger changeovers occurred at the beginning of each semester. Each top management group selected its own successors.

All those who subsequently became managers in the Management Development Corporation started out at the lowest level as staff analysts who worked on project teams just as in the low-structure situation. After an individual had worked as a staff analyst for some time he became a candidate for selection as a project manager. Based on observed performance and anticipated managerial role performance some were promoted to project manager and many were not. Thus, in both the low- and high-structure situations what was involved was picking a good leader based on observation of prior performance in similar situations. In both cases the choices were made by other students; but in one instance they were made in a manner typical of hierarchically structured organizations and in another in a manner typical of organizations with minimal structure. This structural variation was the experimental differentiation.

At a point subsequent to the first move into management as a project manager factors other than observed performance at a lower level assumed increased significance in the promotion process. Seniority began to enter in the promotion process to a sizable degree; so did a personal decision as to whether continuing as a manager in the Management Development Corporation was preferable to other types of learning opportunities. Thus the promotion from staff analyst to project manager differed significantly from those to department manager, general manager, vice president, and president.

The staff analysts derived primarily from the introductory management course which was required of all business students. Participation in the Management Development Corporation was voluntary within the course; students could complete the course requirements in several different ways. However, the grading system was established so as to provide an incentive to work within Management Development Corporation, and a relatively high proportion of the students did in fact include this alternative as one of their activities.

The Miner Sentence Completion Scale was administered in early 1971 to students enrolled in the introductory course in management and to the other more advanced students who had continued as members of the simulated organization. The response rate for the Miner Sentence Completion Scale was over 90%. Three basic samples were established as follows: (*a*) leaders—73 individuals who held management positions in the Management Development Corporation; (*b*) nonleaders—264 individuals who had elected to work in the Management Development Corporation as staff analysts to complete requirements for the introductory management course; (*c*) nonorganization—78 individuals who could have elected to work in the Management Development Corporation as staff analysts, but chose not to do so, and completed the management course requirements in other ways.

It was hypothesized that the Miner Sentence Completion Scale scores in the leader sample would exceed those in both the nonleader and the nonorganization groups. In addition it seemed probable that if any differences existed between the two nonmanagement samples, those who chose the organization should score higher. Reasons for not electing membership in the Management Development Corporation were sufficiently obscure; therefore, this could not be considered a strong hypothesis.

A further differentiation was made within the leader sample on the basis of managerial level. The higher level group consisted of the president, the vice presidents, the general managers, and the department managers—a total of 31 individuals. The lower level group contained 42 project managers, the first-line supervisors. Higher Miner Sentence Completion Scale scores were hypothesized for the higher level managerial sample than for the lower. In addition, comparisons were made between the two leader groups and the two nonleader groups to identify differences that would hold up over two samples of each type. This analysis provided an approximation to replication.

RESULTS

Low-Structure Study

Table 1 contains the mean Miner Sentence Completion Scale scores for the emergent leaders and the total nonleaders. None of the differences attained significance at even the $p < .10$ level. On 5 of the 9 comparisons the total nonleader mean was higher than that for the leaders, in direct contradiction to the hypothesis.

It remained possible that the definite nonleaders within the total nonleader group might have exhibited very low scores. Data on this point are also given in Table 1. However, an examination of the mean Miner scale scores for the three levels of emergent project leadership (leaders, possible leaders, and definite nonleaders) indicated no support for the hypothesis. None of the comparisons among the three samples attained significance at $p < .10$. On 4 of the 9 variables, including the two total score measures, the definite nonleaders had the highest score, in only one instance were they the lowest. The data provided no support for role motivation theory in a context of low organizational structure.

High-Structure Study

Table 2 contains the comparisons involving the various groups. As indicated the major hypothesis was confirmed. The total leaders, who held managerial positions in the simulated organization, had significantly higher means on both the total score measures than both nonleader groups. The subscale comparisons indicated that the degree to which there was the potential for meeting managerial role requirements that involved competitiveness and the performance of routine administrative functions distinguished the organizational leaders from those who did not hold managerial positions.

There was no evidence that those who elected organizational membership at the nonmanagerial level differed at all from those who did not elect membership. However, the low standing out from group score in the nonorganizational sample, as contrasted with that of the managers, did suggest a possible

TABLE 1

MEAN MSCS SCORES FOR EMERGENT PROJECT LEADERS AND NONLEADERS, AND WITHIN THE NONLEADER GROUP FOR POSSIBLE LEADERS AND DEFINITE NONLEADERS (LOW STRUCTURE STUDY)

MSCS measure	Possible leaders ($n = 20$)	Definite nonleaders ($n = 25$)	Total nonleaders ($n = 45$)	Leaders ($n = 27$)
Total score-item	−1.00	− .80	− .89	− .93
Total score-rare	−2.20	−1.68	−1.91	−2.19
Authority figures	− .05	− .12	− .09	.48
Competitive games	.65	.92	.80	.85
Competitive situations	− .95	−1.52	−1.27	−1.70
Masculine role	− .70	− .52	− .60	− .48
Imposing wishes	.20	.00	.09	.33
Standing out from group	.65	.60	.62	.26
Routine functions	− .80	− .16	− .44	− .67

Note. Abbreviation: MSCS = Miner Sentence Completion Scale.

reason for not joining the Management Development Corporation. The Management Development Corporation, compared to other course options available to the students, required more interaction with other students, more group work, and the possibility of making informal and formal presentations to peers, superiors, and faculty. Students who preferred to work independently in more traditional academic roles (e.g., reading, attending lectures, taking exams) may have avoided the Management Development Corporation, and its opportunity to stand out, as an option.

The comparison of managerial levels within the Management Development Corporation failed to yield any significant differences at $p < .10$ or better, although the trend of the data favored the higher level group. Furthermore, as indicated in Table 2, the higher level managers differed from both nonmanagerial groups at $p < .10$ on 5 of the 9 Miner Sentence Completion Scale measures, while the lower level manager group yielded comparable differences on only 2 variables. Overall it seems quite likely that if larger managerial samples had been available, a small, but statistically significant difference between the two levels would have emerged on both total score and both competitiveness measures.

DISCUSSION

One of the most striking findings from the high-structure study was the consistent differentiation between managers and nonmanagers on the measure of motivation to perform routine administrative duties. Experience suggests that this measure reflected a

TABLE 2

MEAN MSCS SCORES FOR APPOINTED ORGANIZATIONAL LEADERS, NONLEADERS, AND THOSE CHOOSING NOT TO JOIN THE ORGANIZATION AND WITHIN THE LEADER GROUP FOR HIGHER LEVEL LEADERS AND LOWER LEVEL LEADERS

MSCS measure	Higher level leaders ($n = 31$)	Lower level leaders ($n = 42$)	Total leaders ($n = 73$)	Non-leaders ($n = 264$)	Non-organization ($n = 78$)	t values					
						Higher level leaders vs.		Lower level leaders vs.		Total leaders vs	
						Non-leaders	Non-organization	Non-leaders	Non-organization	Non-leaders	Non-organization
Total score–item	2.14	.98	1.47	−.76	−1.12	2.98***	2.73***	1.78*	1.76*	3.01***	2.57**
Total score–rare	−1.13	−1.52	−1.35	−2.12	−2.12	2.30**	2.06**	—	—	2.48**	2.03**
Authority figures	.03	−.10	−.04	.09	.36	—	—	—	—	—	—
Competitive games	1.68	1.00	1.29	.85	.85	2.18**	1.89*	—	—	1.83*	—
Competitive situations	−.48	−.98	−.77	−1.36	−1.31	2.38**	2.08**	—	—	2.57**	1.93*
Masculine role	−.33	−.33	−.33	−.64	−.73	—	—	—	—	—	—
Imposing wishes	.55	.55	.55	.55	.49	—	—	—	—	—	—
Standing out from group	.61	.60	.60	.31	−.02	—	1.82*	—	2.21**	—	2.43**
Routine functions	.09	.24	.18	−.55	−.76	2.06**	2.50**	2.93***	3.23***	3.40***	3.62***

Note. t values ($p < .10$) only. MSCS = Miner Sentence Completion Scale.
* $p < .10$.
** $p < .05$.
*** $p < .01$.

type of conscientiousness, or perhaps a mild compulsiveness. In a student-run organization such as the Management Development Corporation this role requirement would appear to be particularly important. There was a strong need to get someone to participate and do the work. Those who exhibited a desire to perform the necessary routine tasks were thus most likely to be selected by the existing managers for new managerial positions. Each group of managers that was promoted to higher levels perceived a need for help in getting its work accomplished and thus tapped a group that exhibited motivation to fill the role requirement to replace the lower level managerial positions that were vacated.

One of the major problems with the Management Development Corporation simulated organization as a teaching device noted by the management faculty at Western Michigan was that a number of students exhibited only minimal involvement. It appears from the present research that the hierarchic leadership selection process operates to provide a solution to this problem at least at the level of the organization, although not necessarily at the level of the individual.

Within the ranks of management there was some evidence that those who enjoyed competitive striving were particularly likely to move up into higher management positions. Actually the managers who reported to the president were selected from among those who volunteered for these positions in preference to some other form of independent or group work. Yet because the simulated organization was staffed and run by students who move from junior to senior standing and then graduate, it departed from most actual organizations in the rapidity with which turnover occurred and openings developed in the managerial ranks, most significantly at the upper levels. The result was that each year a high proportion of the existing project managers had to move up at least one managerial grade level, a much higher proportion than one would find being promoted that fast in most real organizations; seniority per se was a major factor in promotion. As a consequence, the degree of selectivity on the basis of observed performance tended to be much less when upper and lower management were contrasted, than when those promoted into management were contrasted with those who were not. It was presumably, at least in part, because of the artificially inflated turnover rate at the top that the managerial level differences were so small within the simulated organization.

Yet, in spite of this inherent difficulty with the particular simulation used, the evidence in support of role-motivation theory as applied to highly structured administrative or bureaucratic organizations seems quite strong. Total score differences on the Miner Sentence Completion Scale emerged consistently between those appointed to management positions and those not appointed. On the other hand, the research provided no evidence in support of the applicability of the theory in a context of minimal organizational structure.

REFERENCES

GANTZ, B. S., ERICKSON, C. O., & STEPHENSON, R. W. Measuring the motivation to manage in a research and development population. *Proceedings of the 79th Annual Convention, American Psychological Association,* 1971, **6**, 129–130.

GANTZ, B. S., ERICKSON, C. O., & STEPHENSON, R. W. Some determinants of promotion in a research and development population. *Proceedings of the 80th Annual Convention, American Psychological Association,* 1972, **7**, 451–452.

GHISELLI, E. E. *Explorations in managerial talent.* Pacific Palisades, Calif.: Goodyear, 1971.

GHISELLI, E. E., & LODAHL, T. M. Patterns of managerial traits and group effectiveness. *Journal of Abnormal and Social Psychology,* 1958, **57**, 61–66.

MCCLELLAND, D. C., & WINTER, D. G. *Motivating economic achievement.* New York: Free Press, 1969.

MINER, J. B. *Studies in management education.* New York: Springer, 1965. (a)

MINER, J. B. The prediction of managerial and research success. *Personnel Administration,* 1965, **28**(5), 12–16. (b)

MINER, J. B. Changes in student attitudes toward bureaucratic role prescriptions during the 1960s. *Administrative Science Quarterly,* 1971, **16**, 351–364.

MINER, J. B., & DACHLER, H. P. Personnel attitudes and motivation. *Annual Review of Psychology,* 1973, **24**, 379–402.

MINER, J. B., & SMITH, N. R. Managerial talent among undergraduate and graduate business students. *Personnel and Guidance Journal,* 1969, **47**, 995–1000.

(Received October 13, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 38–42

SUPERVISORY BEHAVIOR RELATED TO JOB SUCCESS OF HARD-CORE UNEMPLOYED OVER A TWO-YEAR PERIOD

RICHARD W. BEATTY[1]

Graduate School of Business Administration, University of Colorado

This study examined the impact of two levels of supervision on the job success (job performance and earnings) of 21 hard-core unemployed black women. Over a two-year period supervisory consideration, structure, and insight to intervene in hard-core unemployed interpersonal difficulties were measured. Hard-core unemployed black women's initial perceptions of supervisors were related to early job success and were predictive of future job success. Supervisory consideration was positively associated with job success while supervisory structure was negatively related. It was concluded that attempts to be successful with the hard-core unemployed should not only encourage supervisory styles that are initially supportive of hard-core unemployed job efforts, but should refrain from imposing structure on a culture that may be unaccustomed to external, rigid demands on behavior.

After Bloom and Barry (1967) reported that "hygiene" factors were more important to blacks than whites in entry-level jobs, one hygiene factor, the importance of supervision and the climate supervisors create, has been repeatedly mentioned as critical to the job success of the hard-core unemployed[2] (Ehrenberg, 1967; Goodman, 1969; Hodgson & Brenner, 1968; Lawrie, 1969; Morgan, Blonsky, & Rosen, 1970; National Industrial Conference Board, 1969). In fact, effective supervision appears to be so important that the *Report of the National Advisory Commission on Civil Disorders* (1968) recommended that supervisors be given special training to deal with the hard-core unemployed as one of the basic strategies to reduce urban unemployment rates. This recommendation was not surprising because supervision has long been mentioned as a critical variable in the performance of most employees, regardless of their employment or cultural histories.

Despite the attention given to the supervision of the hard-core unemployed, the relationship between supervisory styles and hard-core unemployed job success has not been investigated over different lengths of time on the job, nor have established instruments been available to assess supervision. To date, the literature that exists in this area can be summarized by noting three variables that continue to be discussed: (*a*) the support the hard-core unemployed receives from a supervisor, (*b*) the amount of close or rigid supervision given the hard-core unemployed, and (*c*) the supervisor's ability to recognize the need to intervene in hard-core unemployed interpersonal difficulties. The need for a generally supportive supervisory climate was demonstrated by Friedlander and Greenberg (1971) and by the National Industrial Conference Board (1970) which found that the hard-core unemployed wanted supervisory support in terms of friendliness, encouragement, and courtesies. This description seems to be no different than the consideration dimension developed for all types of supervisory personnel by the Ohio State Leadership studies (Stogdill & Coons, 1957). Close supervision has been mentioned as generally dissatisfying to the hard-core unemployed

---

[1] Requests for reprints should be sent to Richard W. Beatty, Graduate School of Business Administration, University of Colorado, Boulder, Colorado 80302.

[2] In order to be classified as hard-core unemployed, a person must be a member of a poor family and be unemployed, underemployed, or hindered from seeking work *and* at least one of the following: (*a*) school dropout, (*b*) minority group members, (*c*) under 22 years of age, (*d*) 45 years of age or older, (*e*) handicapped, or (*f*) subject to special obstacles to employment (U.S. Department of Labor, Manpower Administration, *Request for Proposal MA-3,* 1968).

(Goodale, 1973) and described as a means of gaining social control over hard-core unemployed trainees (Krosney, 1966), or as an attempt to control the hard-core unemployed due to differing perceptions of job behavior norms (Triandis & Malpass 1971). Thus it appeared that close supervision of the hard-core unemployed was similar to the definition of *initiating structure,* the second supervisory dimension developed by the Ohio State Leadership studies. The final variable mentioned, defined as social insight, has been cited as the ability to use cooperative behavioral skills to reduce conflict without resorting to hostility (Triandis & Malpass, 1971) or as the ability to intervene and mediate the work environment of the hard-core unemployed, such as in instances of hazing or discrimination (National Industrial Conference Board, 1970).

In addition to reports indicating that the hard-core unemployed were influenced by their immediate supervisors on the above factors, the National Industrial Conferencee Board (1970) reported that supervisory attitudes beyond the first-level (immediate supervisors) were important in successful hard-core unemployed programs, and Nealey and Fiedler (1968) found that second-level supervision (managers) appeared to influence performance more than first-level supervision. Therefore, it appeared as if the level of supervision beyond the hard-core unemployed's immediate supervisor might be important for the hard-core unemployed job success. The purpose of this study, then, was to investigate over different lengths of time hard-core unemployed job success as related to two levels of supervision measured on dimensions of consideration, structure, and social insight. More specifically, four sets of variables were measured: (*a*) the supportive relationship a supervisor provides, (*b*) the closeness of supervision or the degree to which a supervisor defines programmed and controlled hard-core unemployed job behavior, (*c*) sensitivity of the supervisor to intervene in hard-core unemployed interpersonal difficulties, and (*d*) criteria of hard-core job success. These measures were available for all variables after six months and again after two years.

METHOD

Program and Participants

The organization studied was the home office of a large insurance company in the Midwest. The company's president was the metropolitan area chairman of the National Alliance of Businessmen's JOBS program which at least gave the organization's hard-core unemployed effort visible support from the top. The company had over 800 clerical positions for which 41 black women who met hard-core unemployed criteria were hired through funding from the manpower administration of the U.S. Department of Labor. Thirty-one of these women were with the company after six months and 21 remained two years later.[3]

Procedure

To measure the consideration and structure dimensions both the Supervisory Behavior Description and the Leadership Opinion Questionnaire were used.[4] The Supervisory Behavior Description (Fleishman, 1953, 1957a, 1972) and Leadership Opinion Questionnaire (Fleishman, 1957b, 1960) were administered to obtain measures of how the hard-core unemployed perceived the behavior of her supervisor (Supervisory Behavior Description) and the supervisor's attitudes toward supervision (Leadership Opinion Questionnaire). The use of both instruments followed the suggestion of Greenwood and McNamara (1969) who advocated the collection of self-perceived and others-perceived data on supervisory style. Fleishman and Harris (1962) defined *consideration* as:

> behavior indicating mutual trust, respect, and a certain warmth and rapport between the supervisor and his group. This does not mean that this dimension reflects a superficial "pat on the back," "first name calling," kind of human relations behavior. This dimension appears to emphasize a deeper concern for group members' needs and includes such behavior as allowing subordinates more participation in decision making and encouraging more two-way communication [p. 43].

Supervisors' attempts to structure hard-core unemployed job efforts were measured by the initiating structure scale of the Supervisory Behavior Description and Leadership Opinion Questionnaire instruments. Fleishman and Harris (1962) defined *structure* as:

---

[3] The turnover rates for these hard-core unemployed clerical workers were less than that for other clerical workers employed by the company.

[4] All instruments were reworded for administration to the hard-core unemployed. The changes were made by a similar group of hard-core unemployed at the local Concentration Employment Program office. Care was taken to change only language and not intent of a phrase. For example, in some instances "criticizes" was changed to "hassles" and in other instances it was changed to "told you."

> behavior in which the supervisor organizes and defines group activities and his relation to the group. Thus, he defines the role he expects each member to assume, assigns tasks, plans ahead, establishes ways of getting things done, and pushes for production. This dimension seems to emphasize overt attempts to achieve organizational goals [pp. 43–44].

Supervisory sensitivity to intervene in hard-core unemployed interpersonal difficulties, social insight, was measured by Chapin's (1942) Social Insight Scale, which was administered to supervisors. *Social insight* (Chapin, 1942) has been defined as the ability to recognize

> the need to adjust group conflicts or tensions such as known remarks to relax a dangerous intensity, a suggested compromise to attain temporary agreement, a face-saving remark to avoid embarrassment and to preserve status, or to discover the missing part to complete a thought [p. 214].

Gough's (1965) study of the Social Insight Scale indicates that high scorers were seen as perceptive and imaginative, quick to respond to nuances, and quick to sense what others think and feel; conversely, low scorers were seen as inhibited, cautious, bound by convention, and lacking in verve and independence.

The job success variables were measures of hard-core unemployed job performance and weekly earnings. The job performance measure was the company's standard performance-review form used for all clerical employees. After six months and again after two years of supervising the hard-core unemployed, supervisors recorded job-performance data in terms of specific job behavior pertaining to ability and productivity. The company's performance-evaluation form conformed to the recommendations of the American Psychological Association task force for criterion variables (Baxter, Bray, Carp, Dunnette, Ebel, Messick, & Wesman, 1969) by avoiding the use of subjective measures of minority performance (e.g., cooperation) in favor of the actual behavior necessary for successful job performance (e.g., filing and typing for clerical jobs). In addition to reflecting performance with relative objectivity, the form also demonstrated validity when correlated with performance rankings by the company's training coordinator (Kendall's $\tau = .57$, $p < .01$).

The weekly earnings measures were available both at the six-month and two-year intervals, but at six months all hard-core unemployed had the same earnings and thus earnings were not a meaningful criterion. However, after two years there was sufficient variance in earnings for it to be used as a criterion variable. In fact, after two years on the job, the weekly earnings measure became the more important criterion because the hard-core unemployed were placed in different levels of jobs which demanded different skills for which they received different wages. The performance-review form was then less appropriate because it was primarily designed to compare employees in the same job level. Thus, a new employee or a low performer in a high-skill job could receive a low score on the performance-review form while an experienced employee in a low-skill job might receive a high-performance score, although not earn as much as the employee in the high-skill job. The performance-review score was therefore a relative score compared to other employes at the same job level while the weekly earnings measure reflected the worth of the job to the company as well as additional earnings adjustments for performance variance within each job level. The weekly earnings measure thereby became the more important criterion for employees with two years of experience. Job-performance evaluations were then considered the criterion for early hard-core unemployed job success and weekly earnings the criterion for later job success.

RESULTS

No significant relationships were found between hard-core unemployed job success and supervisory consideration, structure, or social insight when supervisors recorded their own scores. The failure of significant relationships to be found might be explained by the fact that the instruments (the Leadership Opinion Questionnaire and the Social Insight Scale) measured only supervisors' self-perceptions and did not measure how the supervisors were actually perceived by the hard-core unemployed as did the Supervisory Behavior Description. Therefore, supervisors might perceive themselves as high in consideration, low in structure, or high in social insight but not be seen the same by their hard-core unemployed employee. Obviously how the hard-core unemployed perceives the attitudes and behavior of a supervisor may have far greater influence on hard-core unemployed job success than does a supervisor's self-perception. Also, in the case of the social insight measure its failure to be related to hard-core unemployed job success not only might be attributed to the hard-core unemployed not feeling that this variable was important, but also to the age of the instrument and/or the extreme difficulty of measuring such a variable, especially where cultural differences may be involved.

Hard-core unemployed perceptions of supervisory behavior did, in fact, demonstrate

TABLE 1

CORRELATIONS BETWEEN FIRST- AND SECOND-LEVEL SUPERVISORY BEHAVIOR DESCRIPTION SCORES AND HARD-CORE UNEMPLOYED JOB SUCCESS

Hard-core unemployed job success	First-level supervision		Second-level supervision	
	Consideration	Structure	Consideration	Structure
6-month supervision with 6-month performance ($N = 31$)	.38*	−.31*	.22	−.11
6-month supervision with 2-year performance ($N = 21$)[a]	.05	.00	−.01	.32
2-year supervision with 2-year performance ($N = 21$)	.22	−.29	−.23	−.02
6-month supervision with 2-year earnings ($N = 21$)[a]	.50**	−.73**	.44*	−.58**
2-year supervision with 2-year earnings ($N = 21$)	.16	−.42*	.16	−.50**

[a] To insure that the reduced sample size had not influenced the significant levels of the six-month measures, the six-month relationships were tested with the reduced sample sizes. All relationships remained significant.

* $p < .05$.

** $p < .01$.

several significant relationships (Table 1). After six months on the job hard-core unemployed perceptions of a first-level supervisor's consideration and structure were concurrently related to hard-core unemployed job performance. Consideration demonstrated a positive relationship with performance, but structure was negatively related.

Two-year hard-core unemployed job-performance measures were neither predicted by the six-month Supervisory Behavior Description consideration and structure scores nor concurrently related to two-year measures of supervisory consideration or structure. However, six-month measures of supervisory consideration and structure were highly predictive of the other job success variable, weekly earnings after two years on the job. Again consideration was positively correlated and structure negatively correlated, but the relationships held for both first- and second-level supervisors. In addition there tended to be a negative concurrent relationship between the hard-core unemployed two-year perceptions of supervisory structure and two-year weekly earnings.

DISCUSSION

The hard-core unemployed's experience with supervisors the first six months on the job appears to be important for initial and future hard-care unemployed job success. The nature of the relationship seems to be that when the hard-core unemployed perceive a supervisor as considerate or supportive they tend to perform more successfully, whereas when a supervisor is perceived as highly defining or structuring work activities, the hard-core unemployed tend to perform less successfully. The positive relationship with consideration seems to be consistent with the findings of Friedlander and Greenberg (1971) and the negative relationship with structure consistent with Goodale (1973).

It was further indicated that the concurrent relationship between supervisory consideration and hard-core unemployed job sucess does not remain after the hard-core unemployed's early experience with the company. This relationship breakdown may be due to the hard-core unemployed's need for supervisory consideration early in their experience with the organization; however, later on other sources of positive reinforcement, such as the work itself or peer support, may explain more of hard-core unemployed job success. Supervisory structure did, however, continue to have a tendency to be concurrently and negatively related to hard-core unemployed job success which may indicate that even after two years on the job the hard-core unemployed may still not respond favorably to supervisory attempts to impose structure on their work efforts.

The finding that six-month consideration and structure scores were both predictive of two-year hard-core unemployed earnings further reinforced the indication that the initial relationship with a supervisor is crucial to hard-core unemployed job success. What may be more important was the finding that not only first-level supervisory behavior but also second-level supervisory behavior was predictive of later hard-core unemployed job success. This finding provides support for the importance of supervisory climate for the hard-core unemployed at all levels of an organization and confirms the results of Morgan, Blonsky, and Rosen (1970).

Thus, this study seems to indicate that programs attempting to be successful in the employment of the hard-core unemployed should, in the initial stages, opt for supervisory styles that are supportive of the hard-core unemployed job efforts, and perhaps permanently refrain from imposing structure on a culture that may be unaccustomed to external, rigid demands on their behavior. The study also demonstrates that when longitudinally investigating the effects of supervision on employee success, it may be useful to (*a*) study levels of supervision beyond the immediate, (*b*) use employee perceptions of supervisory behavior rather than supervisors' self-description, and (*c*) examine criterion alternatives.

REFERENCES

BAXTER, B., BRAY, D. W., CARP, A., DUNNETTE, M., EBEL, R. L., MESSICK, S. J., & WESMAN, A. G. Job testing and the disadvantaged. *American Psychologist*, 1969, **24**, 635–650.

BLOOM, R., & BARRY, J. R. Determinants of work attitudes among Negroes. *Journal of Applied Psychology*, 1967, **51**, 291–294.

CHAPIN, F. Preliminary standardization of a social insight scale. *American Sociological Review*, 1942, 7, 214–228.

EHRENBERG, J. Sub-professionals in programs for the disadvantaged. Paper presented at the annual meeting of American Psychological Association, Washington, D.C., September, 1967.

FLEISHMAN, E. A. The description of supervisory behavior. *Journal of Applied Psychology*, 1953, **37**, 1–6.

FLEISHMAN, E. A. A leader behavior description for industry. In R. M. Stogdill & A. E. Coons (Eds.), *Leadership behavior: Its description and measurement*. Columbus: Ohio State University Bureau of Business Research, 1957. (a)

FLEISHMAN, E. A. The leadership opinion questionnaire. In R. M. Stogdill & A. E. Coons (Eds.), *Leadership behavior: Its description and measurement*. Columbus: Ohio State University Bureau of Business Research, 1957. (b)

FLEISHMAN, E. A. *Manual for leadership opinion questionnaire*. Chicago, Ill.: Science Research Associates, 1960.

FLEISHMAN, E. A. Manual for the supervisory behavior description. Silver Spring, Md.: American Institutes for Research, 1972.

FLEISHMAN, E. A., & HARRIS, E. F. Patterns of leadership behavior related to employee grievances and turnover. *Personnel Psychology*, 1962, **15**, 43–56.

FRIEDLANDER, F., & GREENBERG, S. Effect of job attitudes, training, and organizational climate on performance of the hard-core unemployed. *Journal of Applied Psychology*, 1971, **55**, 287–295.

GOODALE, J. G. Effects of personal background and training on work values of the hard-core unemployed. *Journal of Applied Psychology*, 1973, **57**, 1–9.

GOODMAN, P. S. Hiring, training, and retaining the hard-core. *Industrial Relations*, 1969, **9**, 54–66.

GOUGH, H. A validation study of the Chapin Social Insight Test. *Psychological Reports*, 1965, **71**, 359–361.

GREENWOOD, J. M., & MCNAMARA, W. J. Leadership styles of structure and consideration and managerial success. *Personnel Psychology*, 1969, **22**, 141–152.

HODGSON, J. D., & BRENNER, M. H. Successful experience: Training hard-core unemployed. *Harvard Business Review*, 1968, **46**, 148–156.

KORMAN, A. K. Consideration, "initiating structure" and organizational criteria—A review. *Personnel Psychology*, 1966, **19**, 349–361.

KROSNEY, H. *Beyond welfare: Poverty in the supercity*. New York: Holt, Rhinehart & Winston, 1966.

LAWRIE, J. W. Making it the hardest way. *Psychology Today*, 1969, **3**(6), 29–31, 60.

MORGAN, B. S., BLONSKY, M. R., & ROSEN, H. Employee attitudes toward a hard-core hiring program. *Journal of Applied Psychology*, 1970, **54**, 473–478.

NATIONAL INDUSTRIAL CONFERENCE BOARD. *Education, training, and employment of the disadvantaged*. (Studies in Public Affairs No. 4) New York: Author, 1969.

NATIONAL INDUSTRIAL CONFERENCE BOARD. *Managing programs to employ the disadvantaged*. (Studies in Personnel Policy No. 219) New York: Author, 1970.

NEALEY, S. M., & FIEDLER, F. E. Leadership functions of middle managers. *Psychological Bulletin*, 1968, **40**, 313–329.

Report of the National Advisory Commission on Civil Disorders. Washington, D.C.: U.S. Government Printing Office, 1968.

STOGDILL, R. M., & COONS, A. E. (Eds.) *Leadership behavior: Its description and measurement*. Columbus: Ohio State University Bureau of Business Research, 1957.

TRIANDIS, H. C., & MALPASS, R. S. Studies of black and white interaction in job settings. *Journal of Applied Social Psychology*, 1971, **1**, 101–117.

(Received September 11, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 43-47

A RELATIONSHIP BETWEEN LEADER INTELLIGENCE AND LEADER RATED EFFECTIVENESS

LOUIS S. CSOKA

United States Military Academy

The relationship between a leader's intelligence and his performance has been consistently low. However, these disappointing results can be explained by the contingency model (Fiedler, 1967, 1971) which postulates that the leader's effectiveness depends on his motivational style (least preferred co-worker) and the favorableness of the leadership situation. The approach in this paper conceptualizes intelligence as the leader's ability to integrate his experience and thus provide him with skills to deal effectively with the technical aspects of the tasks and his interpersonal relationships. Four successive studies showed that the relationship between leader intelligence and performance depended on the leader's motivational style (least preferred co-worker), his experience, and the leader-member relations.

A leader's effectiveness has been typically determined either by his personal attributes alone or by some combination of the characteristics of the situation and personal attributes. Campbell, Dunnette, Lawler, and Weick (1970) provided convincing evidence of the limited and partial role played by the leader's personal attributes such as intelligence. The relationship between the leader's intelligence and his group's performance has been generally low and inconsistent. Stogdill (1969), in a review of personal factors associated with leadership, found a positive but rather small relationship (median *r* of .28) between intelligence and leadership. Hollingworth (1926) found that extreme discrepancies between the intelligence of leaders and their followers militate against the successful exercise of leadership. Mann (1959) reported the typical relationship between intelligence and leadership to be between .10 and .25. This finding suggested that the intelligence of the leader is either an insignificant factor or, as this study attempts to postulate, it is poorly used in assigning leaders.

Recent studies by Fiedler (1971) and Csoka and Fiedler (1972) have shown that the leader's experience and training can be interpreted in light of the contingency model. This theory postulates that the effectiveness of the group depends on (*a*) the personality, or more specifically, the motivational system of the leader measured by his least preferred co-worker score, and (*b*) the situational favorableness or the degree to which the situation gives the leader power and influence.

Leadership training can be interpreted as improving the favorableness of the situation by increasing the power and influence of the leader. The contingency model predicts that the additional power and influence is beneficial to certain leaders under some conditions and detrimental under other conditions. Specifically, the model predicts that task-motivated leaders perform best in very favorable and unfavorable situations while relationship-motivated leaders perform best in intermediate situations. It can be said that training leaders for jobs which fall into the favorable situation improves the performance of task-motivated leaders while it decreases the performance of relationship-motivated leaders. It would not be surprising if the leader's intelligence plays a somewhat similar role in changing the favorableness of the situation by modifying the value of experience.

Experience can be viewed as a barrage of

---

[1] This study was conducted under Contract N00014-67-A0103-0012, Office of Naval Research, Department of the Navy, and Contract N0014-67-A0103-0013, Advanced Research Projects Agency, Office of Naval Research (Fred E. Fiedler, principal investigator).

[2] The views expressed in this paper are those of the author and do not reflect thinking of the Department of the Army or of the United States Military Academy.

Requests for reprints should be sent to CPT Louis S. Csoka, Office of Military Psychology and Leadership, United States Military Academy, West Point, N.Y. 10996.

cognitive information that must be properly encoded, structured, and categorized in order to make its use possible. This encoding and structuring requires intelligence. In effect, an interaction of intelligence and experience provides a form of on-the-job training. An intelligent leader can make use of his experience more effectively than a leader with low intelligence. For example, the remark has often been made—sometimes facetiously—that some people do not learn from their experience. An analysis relating leadership experience to changes in performance for relatively intelligent and less intelligent leaders should permit us to infer whether the less intelligent leader is unable to integrate, and therefore, use his leadership experience. In contrast, the intelligent leader might be able to make good use of his leadership experience and thereby increase his power and influence. These analyses could explain the inconsistent results in past research relating intelligence and leadership effectiveness.

Intelligence should be viewed as an intervening variable that modifies the leader's experience and hence the favorableness of the situation. The change in the situational favorability occurs through the modification of the task structure dimension. Even a relatively structured task (objectively defined) should seem initially unstructured for an inexperienced, low-intelligence individual. It appears that not only the objective characteristics of the task should be considered but also the subjective characteristics of the individual, that is, experience and intelligence. The base of knowledge with which the leader enters into the situation must to some degree alter the nature of the task and, therefore, the favorableness of the situation. Therefore, in terms of the contingency model, intelligence is an asset for leaders in some situations but a liability in others. A series of studies have been conducted to test these hypotheses.

Hypothesis 1. In this study intelligence is viewed as that which enables the leader to profit from his experience. Thus, intelligent leaders with long experience should see the task as structured while relatively less intelligent leaders with experience should see the task as less structured. Inexperienced leaders should, of course, see the task as relatively unstructured regardless of their intelligence. Therefore, in terms of the contingency model the following relations between least preferred co-worker and performance can be predicted:

Octant	Group atmosphere	Task structure		Position Power	Predicted relationship
		Experience	Intelligence		
I	High	High	High	High	–
III	High	Low	High/Low	High	–
V	Low	High	High	High	+
VIII	Low	Low	High/Low	Weak	–

Hypothesis 2. Leader intelligence should be related to leader-rated effectiveness only insofar as the leader's motivational style (least preferred co-worker), his experience, and the leader-member relations are taken into account. The complex interactions should reveal the following relationship between leader intelligence and rated performance:

Group atmosphere	Experience	Least preferred co-worker	Rated effectiveness
High	High	High	High intelligence—Lower
High	High	Low	High intelligence—Higher
High	Low	High	High intelligence—Lower
High	Low	Low	High intelligence—Higher
Low	High	High	High intelligence—Higher
Low	High	Low	High intelligence—Lower
Low	Low	High	High intelligence—Higher
Low	Low	Low	High intelligence—Lower

TABLE 1

SPEARMAN RANK-ORDER CORRELATIONS BETWEEN LEAST PREFERRED CO-WORKER AND PERFORMANCE

Leader	Leader-member relations							
	Good				Poor			
	Experience							
	High		Low		High		Low	
	Intelligence							
	Octant I	*n*	Octant III	*n*	Octant V	*n*	Octant VIII	*n*
Field artillery section chiefs	−.54	6	−.52	16	.65*	11	−.52*	22
Navy maintenance supervisors	−.52	10	−.55*	18	.66	6	−.23	21
Company commanders	−.69*	11	−.70**	19	.80*	9	−.53*	19
Company mess sergeants	−.59	9	−.67**	22	.77*	10	−.20	11

Note. Octant I is high intelligence; Octant III is high/low intelligence; Octant V is high intelligence; Octant VIII is high/low intelligence.
* $p < .05$.
** $p < .01$.

METHOD

Subjects

The studies were conducted with subjects in various military situations. Specifically, they were: (*a*) 55 sergeants in charge of field artillery sections, (*b*) 55 navy petty officers in charge of navy aviation maintenance shops, (*c*) 58 unit commanders usually with the rank of captain, and (*d*) 52 sergeants in charge of army mess units. A series of questionnaires was administered to obtain least preferred co-worker scores, the leader-member relations (Group Atmosphere scale), leadership experience, and relative intelligence. It might be noted that all leaders had relatively high position power. The methods and procedures in each of the studies were almost identical. The questionnaires were administered to the subjects in one session for each study. Even though the methodology was similar, the leadership situation for each group of subjects was diverse. While the leaders in three groups were first-line supervisors, the company commanders were second-line or third-line supervisors.

Least-Preferred Co-worker (LPC) scale. Subjects were asked to think of all men with whom they had ever worked and then to describe the one person with whom they could work least well. These descriptions were made on a standard 22-item, bipolar adjective scale. A high score indicated a basic motivation to relate to others while a low score indicated a task-motivation (Fiedler, 1967).

Group Atmosphere scale (GA). This score was obtained by asking individuals to describe on a 10-item, bipolar scale, the atmosphere of their work group. The score reflected the degree to which the leader felt that the group was loyal and supportive of him (McNamara, 1968). The cutting score of high versus low was based on Posthuma's (1970) finding for real-life groups which showed a median group atmosphere score of .655.

Intelligence. A shortened version of the Henmon–Nelson Mental Ability test was administered. The 42-item measure contained both verbal and quantitative scales and had a split-half reliability of .98 and .76 for the verbal and quantitative items, respectively. This score has been shown to be positively correlated with the AGC test scores used by the services to measure intelligence.

Situational favorableness. Situational favorableness was typically measured on the basis of three dimensions: (*a*) leader-member relations, (*b*) task structure, and (*c*) position power. Groups were classified into the upper or lower half on each dimension; each of these dimensions could have been scaled. The three-dimensional system resulted in an ordered set of eight octants scaling the situational favorableness from very favorable to unfavorable. The present studies predict that experience and intelligence interact in affecting the leader's understanding of and ability to handle the task (Csoka & Fielder, 1972). The intelligence and experience interaction modified the task-structure dimension.

Performance measures. Two or three superiors rated each leader's performance with high interrater agreement among subjects. The rating scale consisted of eleven items which were developed in collaboration with the organizations to assess the relative effectiveness of the leaders on critical aspects of the job.

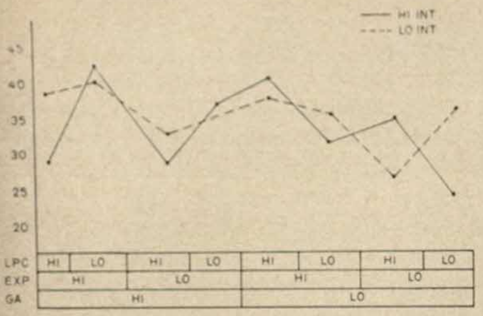


FIGURE 1. Mean performance scores for high- and low-intelligence leaders in Field artillery crews. (Abbreviations: LPC = least preferred co-worker, EXP = experience, GA = group atmosphere score, HI = high, LO = low, INT = intelligence).

RESULTS

It was hypothesized (Hypothesis 1) that the relationship between least preferred co-worker and rated effectiveness should be dependent on the situational favorableness as defined by the leader-member relations, position-power experience, and intelligence. Table 1 shows the results of the four studies. All correlations between least preferred co-worker and performance were in the predicted direction with 10 of the 16 correlations reaching significance. In terms of Fiedler's contingency model, the results showed that the degree to which experience improved the favorableness of the situation was contingent on the intelligence and motivational style (least preferred co-worker) of the leader. When the subjects had low intelligence, they did not seem to gain from their experience. In fact, the less intelligent leaders, under conditions which could be predicted from the contingency model, received higher performance ratings than did more intelligent leaders.

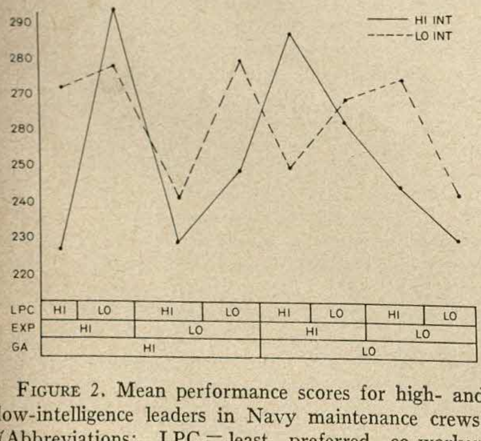


FIGURE 2. Mean performance scores for high- and low-intelligence leaders in Navy maintenance crews. (Abbreviations: LPC = least preferred co-worker, EXP = experience, GA = group atmosphere score, HI = high, LO = low, INT = intelligence).

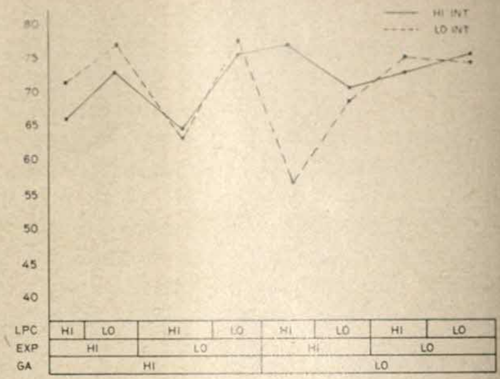


FIGURE 3. Mean performance scores for high- and low-intelligence leaders for Army mess hall stewards. (Abbreviations: LPC = least preferred co-worker, EXP = experience, GA = group atmosphere score, HI = high, LO = low, INT = intelligence).

Hypothesis 2 predicted the nature of the relationship between leader intelligence and rated effectiveness within the operational framework of the contingency model. As Figures 1 through 4 show, the results do support the hypothesis. The relative rated performance of high- and low-intelligence leaders changes depending on their least preferred co-worker, experience, and leader-member relations. The data showed that high intelligence, in and of itself, did not in fact lead to higher-rated performance scores but rather depended upon the complex interaction of several variables.

DISCUSSION

It has been known for some time that leader intelligence does not uniformly increase leadership effectiveness. The present studies enabled one set of conditions under which intelligence *is* related to performance to be specified. Intelligence needs to be viewed as a variable that modifies one's ability to deal with a specific task. This approach did not significantly alter the traditional objective measure of task structure as described in the contingency model. However, it did mod-

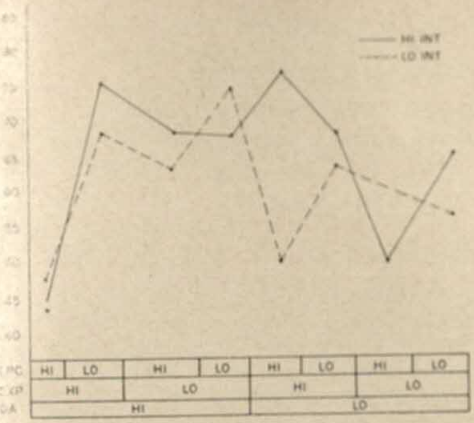


FIGURE 4. Mean performance scores for high- and low-intelligence leaders for Army company commanders. (Abbreviations: LPC = least preferred coworker, EXP = experience, GA = group atmosphere score, HI = high, LO = low, INT = intelligence).

ify the measure to the extent that each individual saw the task as more or less structured depending on his knowledge, which he gained from training and experience. Intelligence was conceptualized as the ability to integrate one's experience. It was found that leaders with high intelligence were not necessarily the higher performers and that when they had very little experience, their intelligence was not beneficial. In effect, the less intelligent leaders could not use their experience and the inexperienced leaders could not use their intelligence in better directing the group.

In accordance with the findings, the interactional effects of the leader's motivational style (least preferred co-worker), experience and leader-member relations showed that high leader intelligence was beneficial to performance only for low least preferred co-worker leaders in favorable situations (high group atmosphere, high experience, strong position power) and for high least preferred co-worker leaders in situations of intermediate favorableness (Low group atmosphere, high experience, strong position power). The findings also showed that less intelligent leaders actually performed better in some conditions than do more intelligent leaders (low group atmosphere, low experience, weak position power).

The results seem to indicate that leader intelligence should be viewed as a moderating variable which affects the usefulness of one's experience. This, in turn, should substantially influence the relative expertise that a leader has in any given situation. Such expertise is essential in the determination of the power and influence of a leader over his members. This power and influence was differentially related to leader effectiveness according to the contingency model. Within the conceptual framework of the model, the four studies shed light on the complexity of any relationship between leader intelligence and performance. The generalizations that can be made from these findings may be somewhat limited due to the similarity of the military subjects in each set of data. At present, the results are meaningful only within the tight operational and theoretical framework of Fiedler's contingency model. Future research should focus on alternate approaches to this relationship in hopes of expanding the meaning of intelligence in leadership research.

REFERENCES

CAMPBELL, J. P., DUNNETTE, M. D., LAWLER, E. E., & WEICK, K. E. *Managerial behavior, performance, and effectiveness.* New York: McGraw-Hill, 1970.

CSOKA, L. S., & FIEDLER, F. E. The effect of military leadership training: A test of the contingency model. *Organizational Behavior and Human Performance,* 1972, **8**, 395–407.

FIEDLER, F. E. *A theory of leadership effectiveness.* New York: McGraw-Hill, 1967.

FIEDLER, F. E. *Leadership.* New York: General Learning Press, 1971.

HOLLINGWORTH, L. S. *Gifted children.* New York: Macmillan, 1926.

KORMAN, A. K. The prediction of managerial performance: A review. *Personnel Psychology,* 1968, **21**, 292–322.

MANN, R. D. A review of the relationship between personality and performance in small groups. *Psychological Bulletin,* 1959, **56**, 241–270.

MCNAMARA, V. D. *Leadership, staff, and school effectiveness.* Doctoral dissertation, University of Alberta, Alberta, Canada, 1968.

POSTHUMA, A. B. Normative data on the least-preferred co-worker (LPC) scale and the group atmosphere (GA) questionnaire. (Tech. Rep. 70-8) Seattle: University of Washington, Organizational Research, 1970.

STOGDILL, R. M. Personal factors associated with leadership: A survey of the literature. In C. A. Gibb (Ed.), *Leadership.* Baltimore, Md.: Penguin Books, 1969.

(Received November 6, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 48–53

CONFLICT AND ATTITUDES TOWARD THE OPPONENT:

AN APPLICATION OF THE COLLINS AND HOYT ATTITUDE CHANGE THEORY TO GROUPS IN INTERORGANIZATIONAL CONFLICT [1]

THOMAS P. CAFFERTY [2] AND SIEGFRIED STREUFERT [3]

Purdue University

Thirty-four four-man teams participated in an experimental simulation of an internation conflict. Each team was subdivided into two decision makers (central roles) and two information handlers (peripheral roles). In keeping with the attitude change theory of Collins and Hoyt (1971), it was predicted that this task allocation, with a possible additional factor of physical separation between subteams, would be a determinant of the players' emergent attitudes toward an opponent team. The results were consistent with the prediction. Teams which engaged in actions with aversive consequences for an opponent held more negative attitudes toward that opponent than teams which (although equal in all other respects) were prevented from *directly* engaging in these actions.

Man, as a social animal, spends much of his time in groups. These groups interact with other groups in a variety of relationships. Some of the relationships reflect cooperation, others competition, and still others neither or both. The research reported in this study deals with small groups in a competitive relationship with one another, and with the differential effects of this competition on attitude formation in members of various subgroups.

Research on intergroup conflict has been primarily concerned with the identification of the causes of conflict and with modes of its resolution (e.g., Deutsch & Krauss, 1960, 1962; Sherif, Harvey, White, Hood, & Sherif, 1961). Relatively little interest has been expressed in the consequences that intergroup conflict has for the group members themselves. These consequences may take a variety of forms. The present research focuses specifically on the attitudinal consequences of participation in conflict. Streufert and Sandler (1971) have shown that the attitudes of decision-making groups toward their opponents tend to be rather negative, reflecting the mirror image perceptions proposed by Bronfenbrenner (1961) and White (1966). It is not known, however, how this negativity toward one's opponents is initially generated. Some relevant information may be provided by reference to studies on the formation and change of attitudes as a function of behavior or commitment to behavior. For instance, a number of studies in the area of dissonance theory have demonstrated that, following the performance of a given behavior, an individual may alter his attitudes to achieve consistency with the behavior he has performed (e.g., Brehm & Cohen, 1962). Kiesler (1971), in a somewhat similar vein, has demonstrated the importance of commitment to behavior as a determinant of the amount and direction of change in task relevant attitudes. Weick (1966) offered evidence that acceptance of an assigned task often involves a kind of attitude change toward the task, which serves to make cognitions consistent with anticipated behavioral requirements. These studies suggest that attitudes may change consequent to performance or commitment to performance of a given task. If we translate this latter proposition into the area of intergroup conflict, it suggests that members of groups who can engage (and have engaged) in competitive actions directed toward an opposing group should hold more negative, that is, task relevant, attitudes toward their opponents than members of groups who are unable to engage (and have not directly engaged) in competitive actions.

---

[1] This research was supported by a contract between the Office of Naval Research, Organizational Effectiveness Research Programs, and Purdue University.

The authors wish to express their thanks to Barry Collins for his helpful comments on the manuscript.

[2] Now at the University of South Carolina.

[3] Requests for reprints should be sent to Siegfried Streufert, Department of Psychological Sciences, Purdue University, West Lafayette, Indiana 47907.

Collins and Hoyt (1971) have recently offered a theory of attitude change that appears relevant to this proposition. These authors suggest that an act must be performed (often obtained via a forced compliance manipulation where the subject believes he can refuse the action, but usually does not) in the presence of perceived choice and aversive consequences; without these conditions attitudes will not be modified. The decision-making groups of Streufert and Sandler (1971), for example, engaged in aggressive decision making toward their opponents. The attitudinal negativity of these groups' views of their opponents may consequently be a reflection of the "attitude change" (here in a negative direction) that, according to the Collins and Hoyt theory might have been expected. Streufert and Sandler's subjects made aggressive decisions; they did have a choice; and they expected aversive consequences for their opponents. Thus, these actions meet the requirements of the theory.

In many settings, for example, in organizations, group tasks are subdivided among several hierarchical or lateral subgroups. Only some of these subgroups might be involved in actual competitive or aggressive activities. If the theory of Collins and Hoyt, and the derivations of other data (previously mentioned), hold, then one may expect differential degrees of negative attitudes toward an opponent among divergent (cooperating) subgroups, depending on the activity they perform, and the perceived choice and consequences. These attitudes may result in the communication and coordination problems found in organizations (cf. Walton, Dutton, & Cafferty, 1969). While decision-making groups that actually engage in competitive acts would be highly negative toward opponents, one might expect that other groups, for example, those involved in the gathering of information, to show much less negativity because they feel less personal responsibility for the consequences of the group decision. The present research design specifically tests this hypothesis.

The comparison between decision-making groups and lateral subgroups that are not involved in the competitive or aggressive decision-making process, however, results in a potential confound. Even though environment and information flow to both kinds of groups may be held constant, instructions to the two kinds of subgroups are necessarily different. The very fact that decision-making groups *can* make decisions, and other lateral subgroups are *not able* to make decisions, might produce differences in behavior that are due to instructional variables and not decision-making activity per se. To test for the Collins and Hoyt theory, control for instructions is required. One way to achieve this control is to introduce group size as an additional variable. If all decision-making groups are specifically instructed to make decisions, and if all information handling groups are specifically instructed to handle information (i.e., not to make decisions), but if half the information handling groups are de facto permitted to engage in decision making, then the design would be adequate. This result can be achieved by placing half the information handling subgroups together with the decision-making subgroups (so that the information handlers can in effect make decisions, have access to decision-making materials, etc.), and by separating the other half of the information-handling subgroups from their decision-making lateral subgroups. Such a manipulation results in a 2 × 2 design (task allocation by group size). Divergent predictions would be made for group size effects, instruction effects, and the effects predicted via the Collins and Hoyt theory. If differences obtained in the data collected with such a design were due to instructions (different for decision makers (DM) than for information handlers (IH) no matter whether separated (2 + 2) or working together (4)), then one would expect that a measure of attitudinal negativity toward the opponents would show that

$$\mathrm{DM}2 + 2 = \mathrm{DM}4 \neq \mathrm{IH}2 + 2 = \mathrm{IH}4.$$

If any obtained differences were due to group size, one would expect that

$$\mathrm{DM4} = \mathrm{IH4} \neq \mathrm{DM2} + 2 = \mathrm{IH2} + 2.$$

If the theory of Collins and Hoyt holds, then information handlers who are placed with the decision makers into a four-man group and consequently are able to participate in the decision-making activity should show more negative attitudes toward their opponents than information handlers who cannot share in the decision-making activity (i.e., those who are placed in a pure two-man information-handling group). Further, the negativity of information handlers who can share in the decision-making activity (in spite of their instructions) should be as great (or nearly as great) as the negativity found in decision-making groups. Finally, if there were no group size effect, then decision makers should be equally negative toward their opponents, no matter whether they work in two-man groups (separated from the information handlers) or in four-man groups (together with the information handlers). In other words, derivation from the Collins and Hoyt theory and the finding of Streufert and Sandler would predict that negativity would differ as follows:

$$\mathrm{IH2} + 2 < \mathrm{IH4} \leq \mathrm{DM4} = \mathrm{DM2} + 2.$$ [4]

In other words, the applicability of the Collins and Hoyt theory to intergroup conflict settings would be supported if groups in the IH2 + 2 condition show less negativity than groups placed in the other three conditions.

As stated previously, the Collins and Hoyt (1971) theory suggests that performance of an act that has aversive consequences for another would be reflected in attitude change. It is not entirely clear whether this attitude change would be a gradual process, producing increasing negativity with repeated performance of acts with aversive consequences, or whether the negative attitudes would be formed as the outcome of one or very few initial actions. (B. Collins, personal communication, 1972, favors the latter interpretation.) To check on the alternatives, repeated measures of attitudes after time intervals of equal length (during which subjects do or do not engage in actions which have aversive consequences for the opponent) are included in this research.

To test the applicability of the Collins and Hoyt (1971) theory, an experimental simulation setting was chosen. The applicability of this research method to organizational problems has been discussed in detail in the handbook chapter of Fromkin and Streufert (in press).

METHOD

One hundred and thirty-six undergraduate volunteers from a large midwestern university participated in the Tactical and Negotiations Game, a complex decision-making environment developed by Streufert and others (Streufert, Castore, & Kliger, 1967; Streufert, Clardy, Driver, Karlins, Schroder, & Suedfeld, 1965; Streufert, Kliger, Castore, & Driver, 1967). In the Tactical and Negotiations Game, subjects were given responsibility to make decisions about military, economic, intelligence, and diplomatic affairs in an experimental simulated small-scale internation conflict with some Vietnam characteristics. Thirty-four four-man teams were formed. The teams spent two hours reading a manual about the game. The purpose of the manual was twofold: (*a*) to expose subjects to the complexities of the conflict with which they later have to deal, and (*b*) to provide two hours of constant preexperimental exposure for all subjects in an attempt to somewhat equalize preexperimental experiences.

After the reading period, subjects in each team were told that they would be assigned in two-man subgroups to one of two different tasks. One subgroup was to function as a decision-making group; the other subgroup was to function as an information-handling group. For 17 of the four-man teams this was the only task manipulation employed. These teams remained operative as four-man teams, even though their assignments to subgroups varied. Typically these teams paid relatively little attention to their assignments, and team members jointly filled information-handling and decision-making functions. For the other 17 teams an additional task manipulation was employed. The subgroups were separated from each other. Each of the two-man subgroups were placed in identical rooms. The rooms in which information-handling subgroups were placed, however, did not contain the forms necessary for decision making. Communication between subgroups was possible via telephone. They were permitted to make as many calls as they wished. Calls were allowed in both directions; information acquisition groups were able to call their decision-making group and vice versa.

---

[4] Further, one might expect that the degree to which information handlers engage in decision-making activity should covary with the degree to which scores for IH4 and DM4 are similar. This proposition is not tested in this study.

Subjects believed that they were playing the Tactical and Negotiations Game against another four-man team. In reality, information received by each team was preprogrammed. All information received by a team was submitted to the team's information-handling subgroup via a mail slot. This subgroup was then responsible for forwarding the information to the team's decision-making subgroup. This procedure was accomplished via the telephone in the subgroups-apart condition. Teams made written decisions that were transmitted to the experimenters via a mail slot. Subjects believed that the outcome of the events in the Tactical and Negotiations Game was greatly influenced by their decisions. (Subjects attributed more than 80% of the causality for the programmed events to decisions of their own and the opposing teams rather than to random events, characteristics of the environment, or arbitrary decisions of the experimenters.)

The experimental simulation was conducted in six 30-minute playing periods, with short intermissions after each period. During the playing periods, the teams received written messages from the experimenters. These messages contained programmed information dealing with the outcomes of the teams' decisions and with various supposed moves by their fictitious opponents. All teams received the same number of messages over the course of the game, but the order in which the programmed messages were submitted was randomized across teams.

Measurement of Dependent Variables

During the intermission following each playing period, the subjects individually filled out a report form consisting of several questions and scales. Among these was a series of ten semantic differential scales on which the subjects were asked to rate their opponents. Three of these were evaluative scales that have been found to be highly intercorrelated in previous research (Streufert, 1965, 1966). When averaged, these scales provide an index of favorable evaluation or liking. The scales used were good–bad, bright–dark, and nice–awful.

RESULTS

The data were analyzed with a three-way mixed design analysis of variance technique containing the following factors: (*a*) group size (two levels, between), (*b*) task assignment (two levels, between), and (*c*) periods (six levels, within). No significance was obtained for the periods main effect ($F = 1.05$) or any of its interactions (all Fs < 1.0). To check on the applicability of the Collins and Hoyt theory on group conflict, the prediction that $\text{IH2} + 2 < \text{IH4} \leq \text{DM4} = \text{DM2} + 2$ was tested. Analysis of variance main effects, interactions, and error term for the Group Size × Task design summed over periods were obtained to permit planned comparisons among the four levels contained in the design.

Check for Confounding Effects

A main effect for task representing task effects for *both* 2 + 2 and four-man groups would have indicated that an instruction confound was present in the data. This main effect did not reach significance ($F = 2.63$, *df* 1/132, *ns*). A main effect for group size that reflected changes in attitudes for *both* IH and DM groups would have suggested a group's size confound that indicated that larger groups form different attitudes toward opponents than do smaller groups. This main effect was significant ($F = 4.03$, *df* 1/132, $p < .05$). However, this difference was reflected only in the IH groups comparison (planned comparison $F = 4.37$, *df* 1/132, $p < .05$). Potential confounds could consequently be excluded from consideration.

Test for Experimental Predictions

Post hoc comparisons based on the analysis of variance error term for the Group Size × Task interaction indicated that the IH2 + 2 groups differed from the combined evaluative ratings obtained in the three other conditions ($F = 6.99$, *df* 1/132, $p < .01$). Single planned comparisons among the sequential elements of the predictions indicated that IH2 + 2 groups produced less attitudinal negativity toward the opponent team than IH4 groups ($F = 4.37$, *df* 1/132, $p < .05$), and that there were no significant differences among the three other groups. The potential difference allowed in the theoretical prediction for comparison of IH4 with DM4 or DM2 + 2 ($F = .23$, $F = .07$, both *ns*) was not obtained. The data are presented in graphic form in Figure 1.

DISCUSSION

The results obtained in this research indicate that negativity toward an opponent in a competitive decision-making task is developed as previously demonstrated by Streufert and Sandler (1971) and as predicted by Brofenbrenner (1961) and White (1966). The negativity in attitudes toward an opponent is modified, however, by the activities in which the group engages. The predictions of Collins

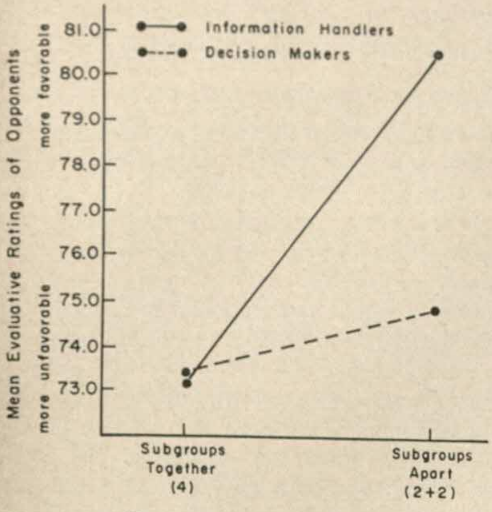


FIGURE 1. Evaluative ratings of opponents by decision makers and information handlers in the Together vs. Separate Subgroups conditions. (Lower scores indicate greater negativity.)

and Hoyt were supported: greater attitude change[5] (here negative attitudes toward opponents) is produced when one makes decisions that have direct aversive consequences for the opponents.[6] Decision makers in this research can be placed into all conceptual categories that Collins and Hoyt propose are necessary for attitude change: (*a*) the decision makers had a choice whether or not to aggress against their opponents; (*b*) all groups did engage in aggressive acts; and (*c*) they viewed their acts to have purposeful negative consequences for their opponents (a necessity in a conflict situation where only one can be the winner, i.e., a situation which subjects view as a zero-sum game). Potential confounds of group size and task relevant instructions were excluded. Further, the differences were obtained after only few decisions with aversive consequences (averaging 1.43 decisions of this kind) had been made. Attitudes later remained stable and were not further affected by later decision making.

It is noteworthy that no differences were obtained for any comparisons between IH 4 groups and both kinds of decision-making groups. It appears that the ability to engage in decision-making activity with the indicated choice and consequence effects was a sufficiently strong manipulation. Effects of instructions by the experimenters did not diminish the effect of the placement of the IH 4 groups into a situation where they were able to contribute to the decision-making activity of their groups. (A check on the required signatures on decision forms indicates that IH persons in these groups actually made 41% of the decisions. In contrast, checks on telephone conversations showed that decision suggestions by IH 2 + 2 groups occurred very rarely.)

The results of this research are encouraging. The theoretical predictions from attitude theories have generally had rather limited implications for other areas of social and organizational psychology. More seriously, opposing theoretical predictions of various writers have been supported by parallel research designs using often similar variables at diverse levels (cf., Collins & Hoyt, 1971). Contradictory findings of this kind often suggest that the phenomenon under study is multidimensional rather than unidimensional in nature (cf., Streufert & Fromkin, 1972), and consequently a number of predictors or conditions may be necessary to produce an effect reliably. Such multiple predictors are contained in the Collins and Hoyt theory. The successful application of the theory to a complex (itself multidimensional) group conflict setting suggests that the view of Collins and Hoyt may prove robust and might be potentially useful as an initial integrative concept for the use of attitudes in a number of research areas in social and organizational psychology.

---

[5] In the Tactical and Negotiations game attitudinal negativity is produced after the participant reads a manual that contains bias against the opponent, and after he participates in a series of economic and military battles with the opponent. The degree of resulting negativity (attitude change) toward the opponent can vary, however.

[6] One might argue that the lesser negativity of IH $2 + 2$ groups could be a reflection of an alienation phenomenon. Such alienation, however, would express itself as well in other ratings obtained from such groups. No such differences (e.g., attitudes toward their own group or in ratings of the degree of success and failure either team had experienced at any point in time were obtained, suggesting that the finding is due to differential attitudes, rather than alienation. However, similar results ($p < .05$) were obtained when subjects were asked to rate the competence of the opposing team on a 7-point rating scale. Again, IH $2 + 2$ teams viewed the opposing team as more competent than any of the other three groups. Since competence ratings tend to be positively correlated with evaluative attitudes (Cafferty & Streufert, 1971); this finding supports the interpretation of data in line with the predictions of Collins and Hoyt.

REFERENCES

BREHM, J. W., & COHEN, A. R. *Explorations in cognitive dissonance.* New York: Wiley, 1962.

BRONFENBRENNER, U. The mirror image in Soviet–American relations: A social psychologist's report. *Journal of Social Issues,* 1961, **17**, 45–56.

CAFFERTY, T. P., & STREUFERT, S. *The group viewed from the vantage point of the individual: Effects of environmental ambiguity on evaluative attitude, perceived competence and perceived influence of central and peripheral group members.* (ONR Tech. Rep. No. 38) Lafayette, Ind.: Purdue University, 1971.

COLLINS, B. E., & HOYT, M. F. *Magnitude of inducement, consequences, and responsibility–choice: A theoretical and empirical integration of the forced compliance literature.* Unpublished manuscript, University of Wisconsin, 1971.

DEUTSCH, M., & KRAUSS, R. M. The effect of threat on interpersonal bargaining. *Journal of Abnormal and Social Psychology,* 1960, **61**, 181–189.

DEUTSCH, M., & KRAUSS, R. M. Studies of interpersonal bargaining. *Journal of Conflict Resolution,* 1962, **6**, 52–76.

FROMKIN, H. L., & STREUFERT, S. Laboratory experimentation. In M. Dunnette (Ed.), *Handbook of organizational and industrial psychology.* Chicago: Rand McNally, in press.

KIESLER, C. A. *The psychology of commitment.* New York: Academic Press, 1971.

SHERIF, M., HARVEY, O. J., WHITE, B. J., HOOD, W. R., & SHERIF, C. W. *Intergroup conflict and cooperation: The Robbers Cave experiment.* Norman, Oklahoma: University Book Exchange, 1961.

STREUFERT, S. Communicator importance and interpersonal attitudes toward conforming and deviant group members. *Journal of Personality and Social Psychology,* 1965, **2**, 242–246.

STREUFERT, S. Conceptual structure, communicator importance, and interpersonal attitudes toward deviant and conforming group members. *Journal of Personality and Social Psychology,* 1966, **4**, 100–103.

STREUFERT, S., CASTORE, C. H., & KLIGER, S. C. *A tactical and negotiations game: Rationale, method, and analysis.* (ONR Tech. Rep.) Rutgers University, 1967.

STREUFERT, S., CLARDY, M. A., DRIVER, M. J., KARLINS, M., SCHRODER, H. M., & SUEDFELD, P. A tactical game for the analysis of complex decision making in individuals and groups. *Psychological Reports,* 1965, **17**, 723–729.

STREUFERT, S., & FROMKIN, H. L. Complexity and social influence. In J. Tedeschi (Ed.), *Social influence processes.* Chicago: Aldine Atherton, Inc., 1972.

STREUFERT, S., KLIGER, S. C., CASTORE, C. H., & DRIVER, M. J. A tactical and negotiations game for analysis of integration across decision areas. *Psychological Reports,* 1967, **20**, 155–157.

STREUFERT, S., & SANDLER, S. I. A laboratory test of the mirror image hypothesis. *Journal of Applied Social Psychology,* 1971, **1**, 378–397.

WALTON, R. E., DUTTON, J. M., & CAFFERTY, T. P. Organizational context and interdepartmental conflict. *Administrative Science Quarterly,* 1969, **14**, 522–542.

WEICK, K. E. Task acceptance dilemmas: A site for research on cognition. In S. Feldman (Ed.), *Cognitive consistency.* New York: Academic Press, 1966.

WHITE, R. K. Images in the context of international conflict: Soviet perceptions of the U.S. and the USSR. In H. C. Kelman (Ed.), *International behavior.* New York: Holt, Rinehart, & Winston, 1966.

(Received September 27, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 54–60

EFFECT OF WORKER ORIENTATION DIFFERENCES ON JOB ATTITUDE CORRELATES

DONALD A. WOOD[1]

Indiana University

Correlations between various worker attitudes and job motivation, performance, and absenteeism were examined for 290 skilled and semiskilled papermakers. The hypothesis that significant relationships occur more frequently for those employees least job involved was supported, inasmuch as they accounted for 84% of the significant correlations. Further analyses disclosed twice as many associations for skilled as for semiskilled employees. Thus, highly involved employees, more intrinsically oriented toward their job, did not manifest satisfactions commensurate with company evaluations of performance; they depended more on intrinsic rewards. Those employees more detached from the job itself were more extrinsic in orientation and experienced gratifications more in line with company performance assessments due to their greater dependence on extrinsic rewards.

It has been surprisingly difficult to demonstrate appreciable relationships between work performances and the satisfactions supposedly inherent in them (Ronan, 1970). Despite frequent disappointments, research in this area continues. These efforts seem to be based on the unacceptable premise that although performance and attitude are often unlike, they should not be. Even when similarities are found, they are often barely significant (Vroom, 1964).

This study examines the correlational issue further by directing attention to the job incumbent and his work orientation. The theoretical emphasis on intrinsic–extrinsic factors of job satisfaction (Herzberg, Mausner, & Snyderman, 1959) and the proposition that intrinsic and extrinsic rewards serve as satisfaction–performance moderators (Lawler & Porter, 1967) have prompted much needed research on some vital issues. However, the conceptual confusion that results from the finding that job satsfaction can originate from either job-related (intrinsic) or situational (extrinsic) features (Burke, 1966) and the inability to operationalize intrinsic rewards creates certain doubts as to whether such distinctions can and should be pursued further.

It is important to focus on the assessment of the intrinsic–extrinsic orientation of the employee himself. Therefore, this study examined the neglected question of the impact of worker orientation differences on satisfaction–performance relationships. Such associations may well be influenced by the extent to which the worker is psychologically committed to work in general, that is, either intrinsically oriented and highly identified with work (high job involvement) or extrinsically oriented and not psychologically dependent on the work itself (low job involvement). The intrinsic and extrinsic nature of high- versus low-involved workers, respectively, has been described elsewhere (Reitz, Groff, Ruiz, & Terry, 1973).

Saleh and Grygier (1969) described job orientation differences in detail. It was found that those intrinsically oriented toward their work were more independent, confident, flexible, less dependent on guidance or reassurance, and likely to reject conventional standards more than their extrinsically oriented counterparts. It would follow that intrinsically oriented or highly involved workers would be more dependent on their own standards to determine appropriate performance levels and less concerned with organizational prescriptions of what performance should be. Vroom (1964) would support this contention by noting that "some rank and file workers may be relatively indifferent to information from supervisors or others regarding their job performance [pp. 146–147]." In addition, Vroom (1962) has concluded that some employees may derive job satisfaction from the performance of tasks independent of the externally mediated consequences of that performance.

---

[1] The author would like to express his appreciation to Dennis W. Organ and H. J. Reitz for their insightful comments on an earlier version of this article.

Requests for reprints should be sent to Donald A. Wood, Graduate School of Business, Room 650, Indiana University, Bloomington, Indiana 47401.

Thus, differences in job orientation or involvement would be expected to promote differences in sensitivity to performance information. Intrinsically oriented or highly involved incumbents should be most reliant on their own performance standards to determine their work gratifications while those less committed to work per se, in the absence of any personalized standards, should be more dependent on external or "alternative" criteria as defined by the organization.

It is hypothesized in this study that performance as evaluated by the organization does not correlate with individual satisfaction levels for highly involved workers. Since the self-concept is not as work related for subjects more job detached, job satisfaction should be more related to the firm's appraisal standards. Thus, it is hypothesized that least involved employees demonstrate significant satisfaction–performance relationships.

METHOD

Sample

The initial sample consisted of 352 semiskilled and skilled male papermakers in a large unionized paper mill plant located in the southeastern United States. The average employee had been employed at the plant for 13 years and was 39 years old. To insure the greatest validity and stability in involvement, satisfaction, and performance assessments, employees who had not been on their present job at least three months and employed by the plant for at least six months were eliminated from the study. This elimination resulted in a loss of 62 employees and yielded 110 semiskilled and 180 skilled workers or a total of 290 in the final sample. Because each skill level was amply represented and differed so in hourly wages received, plant seniority (progression lines to skilled jobs were prescribed contractually on the basis of both merit and seniority) and the complexity of work performed, it was decided to examine the moderating effects of involvement separately for the two skill levels as well as the total sample.

Questionnaire

A 31-item questionnaire was administered on company time to all incumbents during a two-week period. Twenty Likert-scaled items were concerned with the construct of job involvement (Lodahl & Kejner, 1965) and were designed to measure the nature of job orientation by examining the extent to which work in any way is seen to affect one's own sense of personal worth. Because of the well-established multidimensional nature of this construct, as demonstrated by Lodahl and Kejner, a principal components analysis (Dixon, 1968) that employed orthogonal rotation and multiple correlation diagonal estimates was performed. The solution that best represented all 20 items would ultimately provide a series of involvement dimension scores to be used for subsequent sample subgrouping.

The remaining 11 items assessed on a 5-point scale the degree to which incumbents were either satisfied (+5) or dissatisfied (+1) with their present job. Specifically, items pertained to attitudes toward co-workers, wages, job security, recognition, supervision, company policy, advancement, fringe benefits, work itself, job safety, and overall job feelings (for factor-analytic derivations of these dimensions see Wood & LeBold, 1970).

The greatest anonymity possible was preserved for each respondent by requiring only his clock card number on the questionnaire form and by assuring each subject that forms would be mailed to the author immediately following each survey session under union steward supervision. This procedure made it impossible for anyone within the plant to examine any individual questionnaire or items and hopefully assured each respondent that complete honesty to the inventory questions would be possible.

Behavioral Indexes

Three different job behaviors to be later correlated with job attitudes were selected on the basis of their performance diversity and the high weight given them by the company in evaluating personnel. These included supervisory ratings, motivation, and absenteeism.

Supervisory ratings. Paired comparison ratings were secured from first-, second-, and third-line supervision. All possible employee pairs within an occupation under a rater's jurisdiction comprised an evaluation group (unskilled subjects excluded). If numbers did not permit, occupations judged to be similar in duties and skills required were combined into one rating group. Each supervisor was instructed to consider which of the two employees in a given pair was currently performing most satisfactorily on his present job and then to indicate his preference by placing a check after the name. This checking process continued until every employee had been compared and rated with every other worker in his group.

The number of times each employee had been rated first by each rater was converted to a standardized performance index using T scale values (for conversion table see Lawshe & Balma, 1966). The average of these indexes across the three supervisory lines provided the performance score which was correlated with the job satisfaction items.

Motivation. Following the comparison ratings on job performance, supervisors were given a 5-point motivation scale for each of their semiskilled or skilled subordinates. The scale values were (*a*) too uninterested in job and lacks motivation, (*b*) somewhat indifferent toward job, (*c*) some job interest, but not overly motivated, (*d*) actively interested in job, and (*e*) very interested in job and highly motivated. The rater was instructed not to consider actual performance or ability to perform, but only the desire or willingness to perform. Following this determination, the rater merely checked the most appropriate scale number for each subject.

Absenteeism. The third measure of work behavior concerned the frequency of absences for each worker, not the total number of work days missed, during a period of 18 months just prior to this study. In other words, the consecutive absences were counted only as one absence, since the frequency of absence and not the duration was considered more important.

Preliminary Analyses

The 33 unmoderated correlation coefficients between the 11 satisfaction items and the three behavioral indexes for the entire sample disclosed only one exceeding the .20 level (between satisfaction with supervision and job motivation, $r = .23$, $p < .01$). The median correlation of only .07 demonstrated the lack of any meaningful affect–behavior associations when the total sample was examined in a single analysis. The independence between job satisfaction and job involvement was supported by a median item intercorrelation across constructs of only .07 (*ns*). This finding would support the factoral independence of these concepts reported by Lawler and Hall (1970). Median item intercorrelations for job satisfaction and involvement were .33 ($p < .001$) and .16 ($p < .01$), respectively.

The components analysis of the involvement intercorrelation matrix was conducted to determine the extent of subgrouping required. This analysis revealed five factors which represented all 20 involvement items. These included (*a*) work attraction—seeing work as having positive appeal and preparing oneself before each shift begins (*b*) failure sensitization—being bothered by work mistakes and depressed by job failure; (*c*) work commitment—choosing work and its related responsibilities as more important than other activities; (*d*) job preeminence—job provides the major source of gratification when compared to all other activities; (*e*) work identification—perceiving the importance of being identified with work and evaluating others worth on that basis.

Five component scores were computed for each employee by summing across the raw responses (+5 for strongly agree to +1 for strongly disagree) on all items loading above .30 on each dimension. Factor score correlations ranged from .17 to [illegible]. Each score range for each involvement dimension was then divided at the median value for the total sample. This division produced a low-involvement group (below median) and a high-involvement group (above median) on each of the five factors. Next, correlation coefficients were estimated between the 11 satisfaction items and the three performance measures within each of the resulting 10 subgroups.

RESULTS

The dramatic effect of differential job involvement on satisfaction–performance relationships is shown in Table 1. Sixty-four of the 76 significant moderated correlation coefficients appeared for those with low factor scores and offered strong support for the contention that workers low in job involvement manifest many more associations between job satisfaction and company assessments of per-

TABLE 1

FREQUENCIES OF SIGNIFICANT ATTITUDE–BEHAVIOR CORRELATIONS ($p \leq .05$) FOR FIVE INVOLVEMENT DIMENSIONS

Job involvement dimensions	Low involvement			High involvement		
	Motivation	Performance	Absences	Motivation	Performance	Absences
Work attraction	5 (.20*)	1 (.16*)	4 (−.19*)	2 (.23**)	1 (.18*)	0
Failure sensitization	8 (.21*)	0	8 (−.18*)	3 (.18*)	0	0
Work commitment	5 (.17*)	0	6 (−.23**)	2 (.18*)	0	0
Job preeminence	7 (.21**)	3 (.16*)	3 (−.17*)	0	1 (.22*)	0
Work identification	5 (.34**)	2 (.26**)	7 (−.30**)	2 (.19**)	1 (.17)*	0
Subtotals	30	6	28	9	3	0

Note. Significant correlation overall range was from .16 ($p < .05$) to .35 ($p < .001$). Average or actual correlations appear in arentheses.

* $p < .05$.

** $p < .01$.

TABLE 2

ATTITUDE–BEHAVIOR CORRELATIONS FOR LOW– VS HIGH–INVOLVEMENT GROUPS BY INVOLVEMENT FACTOR AND SKILL LEVEL

Satisfaction with	Work attraction						Failure sensitization						Work commitment						Job preeminence						Work identification					
	Low			High			Low			High			Low			High			Low			High			Low			High		
	M	P	A	M	P	A	M	P	A	M	P	A	M	P	A	M	P	A	M	P	A	M	P	A	M	P	A	M	P	A
Co-workers			_−46***_												−23*			22*												
Wages	_29*_														−30**			21*							21*					
Security							21*								−27*			22*	29**	23*				20*						
Recognition	36*** _34*_	20*	_−35*_				29**	_28*_			20*		29*		−25*				34**	33** _36**_	_−27*_				29**	26**				
Supervision	34*** _37**_	23*	−23*				35***	26*	−36***	20*			34**	32**	−27*		_31*_	_−30*_	26*	23*	−23*	27**			30**	22*	−30**			
Company			_−46***_	_35**_			21* _35*_	_33*_	_−34*_				_28*_						24* _28*_	_31*_	−29**						−19*	_48***_		
Advancement																			27*	28*					20*			_32*_		
Fringe benefits				31** _31*_	_27*_		_35*_		_−32*_	21*									_36**_	_34**_					19*			_40**_		
Work itself			_−38**_																				_27*_							
Safety				_28*_			31**	24*					22*						_38**_	_28*_			_31*_							
Overall job			−19* _−47***_	24*			28**	27*	25*						−42***				27* _31*_	33**	−23*		_34*_		26*	18*	−25** _−24*_			

Note. Semiskilled correlations are italicized; skilled correlations are not. M = motivation, P = production, A = absences. Low involvement on a factor means below median score; high involvement means above median score. All decimal points for correlations are omitted.

* $p < .05$.
** $p < .01$.
*** $p < .001$.

formance than do those highly job involved. This trend held for all five dimensions examined.

One of the more striking findings in Table 1 is that analyses in the low-involvement groups produced 28 significant correlations between satisfactions and absenteeism (out of 55), while none could be found for those highly involved. This result would certainly suggest that where such relationships are minimal or nonexistent in a total sample sense, to seek out only those subjects expressing low affinity toward work may well provide a most fruitful means for their enhancement.

The significant moderated correlations presented in Table 2 by involvement factor and skill level ranged from .18 to .48. The advantage of subgrouping in such analyses can be better appreciated when it is recalled that only one nonmoderated correlation exceeded .20.

The most prominent correlates on the attitude side concerned job recognition, supervision, and overall satisfaction, with the skilled groups accounting for the vast majority of such associations (see Items 4, 5, and 11 totals). Satisfactions with co-workers, wages, advancement, and the work itself were particularly limited in relatability for both skilled and semiskilled subjects.

Probably the most notable finding detailed by the table, however, was the marked superiority in significance frequencies for those with low- as opposed to high-factor scores. Low job preeminence was especially productive of such frequencies followed by low failure sensitization status. It should also be noted that except for work attraction, frequencies for skilled employees far exceeded those for semiskilled within the low groups.

Table 3 is offered as a summary aid for data in Table 2. Seventy-eight percent of the significant relationships appeared under low job attachment. Also, it is well confirmed in this study that satisfactions were related to behavior much more often for skilled than for semiskilled employees (64% of the correlations).

Likert (1961) has suggested that satisfaction and performance should be more positively related as skill level increases, a premise which Vroom's (1964) research review could not convincingly verify. Further doubt is offered by the present study, since the median significant correlation for semiskilled was actually higher than for skilled—from .33 ($p < .01$) to .26 ($p < .01$). The frequency of significant correlations was much higher for skilled employees, but their magnitudes were not. Since the more able would tend to gravitate toward the more skilled jobs, this finding would seem congruent with Carlson's (1969) that attitudes related to performance most often for the more able incumbents.

The correlational prominence for those least involved and most skilled held for each of the three job behaviors examined (see Table 2). For job motivation, more than two thirds of the correlations came from skilled ranks and five out of six were in low-involvement groups. For production, 6 correlations in 10 were in skilled groups and 4 in 5 appeared in low-attachment groups. For motivation, 2 out of 3 correlations were for

TABLE 3

FREQUENCY SUMMARY TABLE FOR TABLE 2 CORRELATIONS

Employee	Motivation			Absences			Production			Grand Total		
	LI	HI	Total	LI	HI	Total	LI	HI	Total	LI	HI	Total
Skilled	23	5	28	17	4 (All positive)	21	14	1	15	54	10	64
Semiskilled	10	6	16	8	1	9	6	5	11	24	12	36
Total	33	11	44	25	5	30	20	6	26	78	22	100

Note. LI = low involvement; HI = high involvement.

skilled subjects and 3 in 4 came from the least involved subgroups.

DISCUSSION

It may be concluded from this study that employees least involved with work exhibit much more similarity between their job feelings and the organization's evaluation of their performance than is the case for those expressing a more intrinsic or involved orientation. In the former instance, high job satisfaction is often associated with high motivation, high performance, and low absenteeism. For the latter group, this result is rarely the case; behaviors and attitudes show little systematic correspondence.

Surprisingly, this correlational phenomenon held for all five factors employed in the moderation, despite their distinctiveness. Thus, it would seem that low sensitivity to work generally, irrespective of any specific aspect, serves to accentuate the correlations.

In light of the findings of this study, it seems important to separate perceived performance as determined by self-appraisals from company assessments of acceptability. Strauss (1966) reported distinct differences between self-ratings of performance and assessments by peers and supervisors and demonstrated the importance of considering *which* performance measure is to be employed in correlational questions regarding attitudes.

The whole issue of performance assessment is closely linked with the question of whether rewards are intrinsically (self) administered or extrinsically (company) administered. The problem has been one of operationally distinguishing between workers more sensitive to either reward type.

Worker involvement would almost by definition provide a solution. One involved with or intrinsically oriented toward work would perceive a stronger and more direct relationship between his self-concept and the work performed than one more extrinsically disposed (Lawler & Hall, 1970). It would follow that highly involved incumbents placing more worth on the job itself would be more capable of rewarding themselves intrinsically. In turn, low-involvement employees who do not identify with the work as closely would be more sensitive to externally mediated rewards from the organization. Thus, the differences in job orientation could well be used in classifying those who are capable of self-administration of rewards from those who are not.

If those highly involved are more prone to self-mediated reinforcements, then performance can be rewarded irrespective of its "goodness" or "badness" in any external sense. Consequently, if an organizationally defined standard of job performance, as used in this study, is compared with their satisfactions, little relationship should be evidenced because of the possibility of job satisfaction occurring at many levels of performance, high or low.

Continuing the same reasoning, if low-involved subjects are more sensitized to external rewards and if reasonable contingencies exist between such rewards and company-defined performance standards, then these subjects should experience satisfaction through these rewards more in line with prescribed performance levels. Since they would not likely define work standards for themselves, satisfactions would be more dependent on how the firm evaluates and approves their work behavior.

To link reward theory with varying sensitivities of workers due to differential job involvement seems to offer a reasonable theoretical basis to explain the results obtained in this study. Consequently, it would appear that the Lawler and Porter (1967) model that depicts equitable rewards intervening between performance and satisfaction could be qualified to accommodate differences in job orientation and performance assessment. For example, for low-involvement subjects, job performance evaluated by the organization (JP_{org}) leads to extrinsic rewards (ER) which in turn leads to job satisfaction (JS) or

$$JP_{org} \longrightarrow ER \longrightarrow JS$$

The more appropriate model for those highly involved would seem to be

$$JP_{org} \text{—/—/—/}\!\!\longrightarrow IR \longrightarrow JS$$

The problem for the latter subgroup seems to be that the organization's appraisal of performance does not give rise to the intrinsic rewards on which this subgroup was most dependent for work gratifications. As a result, performance so measured does not relate to satisfactions. Further research is needed to determine if self-evaluated performance would provide the necessary intrinsic rewards demanded by such subjects which in turn would be reflected in commensurate satisfaction levels.

In addition, longitudinal studies of how workers become identified with work are required. It may well be that highly involved subjects have a history of positive reinforcement procured through prior work pursuits. If so, they may be more resistant to organizational rewarding strategies, since they are more capable of initiating and perpetuating their own reinforcement system. On the contrary, those least involved, who may lack such a history, may be more receptive to current organizational rewards, as this study strongly suggests. Worker orientation differences toward work may well determine whether the organization indeed has any power to reward and satisfy effectively.

REFERENCES

BURKE, R. J. Are Herzberg's motivators and hygienes unidimensional? *Journal of Applied Psychology*, 1966, **50**, 317–321.

CARLSON, R. Degree of job-fit as a moderator of the relationship between job performance and job satisfaction. *Personnel Psychology*, 1969, **22**, 159–168.

DIXON, W. J. (Ed.) *BMD biomedical computer programs*. Berkeley and Los Angeles: University of California Press, 1968.

HERZBERG, F., MAUSNER, B., & SNYDERMAN, B. *The motivation to work*. New York: Wiley, 1959.

LAWLER, E. E., & HALL, D. T. Relationship of job characteristics to job involvement, satisfaction and intrinsic motivation. *Journal of Applied Psychology*, 1970, **54**, 305–312.

LAWLER, E. E., & PORTER, L. W. The effect of performance on job satisfaction. *Industrial Relations*, 1967, **7**, 20–28.

LAWSHE, C. H., & BALMA, M. J. *Principles of personnel testing*. New York: McGraw-Hill, 1966.

LIKERT, R. *New patterns of management*. New York: McGraw-Hill, 1961.

LODAHL, T. M., & KEJNER, M. The definition and measurement of job involvement. *Journal of Applied Psychology*, 1965, **49**, 24–33.

REITZ, H. J., GROFF, G. K., RUIZ, A. E., & PERRY, C. M. Similarities and differences among Mexican workers in attitudes important to worker motivation. (Indiana Business Report. No. 47.) Bloomington: Indiana University, School of Business Division of Research, 1973.

RONAN, W. W. Individual and situational variables relating to job satisfaction. *Journal of Applied Psychology*, 1970, **54**(1, Pt. 2).

SALEH, S. D., & GRYGIER, T. G. Psychodynamics of intrinsic and extrinsic job orientation. *Journal of Applied Psychology*, 1969, **53**, 446–450.

STRAUSS, P. Psychology of the scientist: XIX. Job satisfaction and productivity of engineers and scientists. *Perceptual and Motor Skills*, 1966, **23**, 471–476.

VROOM, V. Ego-involvement, job satisfaction, and job performance. *Personnel Psychology*, 1962, **15**, 159–177.

VROOM, V. *Work and motivation*. New York: Wiley, 1964.

WOOD, D. A., & LEBOLD, W. K. The multivariate nature of professional job satisfaction. *Personnel Psychology*, 1970, **23**, 173–189.

(Received September 11, 1972)

Journal of Applied Psychology
1974, Vol. [illegible], No. 1, 61-68

A GUTTMAN APPROACH TO THE DEVELOPMENT OF BEHAVIORALLY BASED RATING SCALES FOR SYSTEMS ANALYSTS AND PROGRAMMER/ANALYSTS[1]

RICHARD D. ARVEY[2]
University of Tennessee
and
Applied Psychological Systems, Inc.
Knoxville, Tennesesee

AND

JOSEPH C. HOYLE
The National Cash Register Company
Dayton, Ohio

Behaviorally based rating scales, for programmer/analysts and systems analysts, which exhibited adequate Guttman scaling properties, were developed. These scales also demonstrated reasonably good convergent and discriminant validity. An attempt was made to identify relatively poor raters using Guttman scaling procedures. However, only slight evidence showed that supervisors who rate poorly on one dimension also rate poorly on other dimensions, or that supervisors who rate one individual poorly also rate another individual poorly on the same dimension.

Developing adequate measures of job performance has certainly been one of the more challenging tasks in contemporary industrial psychology. This task becomes somewhat more difficult when the jobs are of a technical and complex nature, as for example, computer personnel jobs.

The development of behaviorally based rating scales based on the procedure of scaled expectations (Smith & Kendall, 1963) and variants of this method (Campbell, Dunnette, Lawler, & Weick, 1970; Folgi, Hulin, & Blood, 1971) have seemingly resulted in reliable and valid job criteria that help to minimize rater-biasing factors. Briefly, behaviorally based rating scales reflect the major relevant dimensions of job performance where specific behavioral incidents serve as anchor points along the rating scales which represent these dimensions (Smith & Kendall, 1963; Folgi et al., 1971; Campbell et al., 1970). The behavioral anchors are ordered or scaled so that the top level of the rating continuum is anchored by those incidents that represent highly effective behavior within that domain of job performance. Similarly, the low end of the continuum is anchored by incidents of highly ineffective job behavior.

Guttman scaling procedures (Edwards, 1957) can be used to help evaluate behaviorally based rating scales. This technique can be employed to determine whether the behavioral incidents representing anchor points are properly scaled. For example, Schultz and Siegel (1961) presented data which indicated that checklists for evaluating task performance in several related job specialties met Guttman scalability requirements. Hellervik, Dunnette, and Arvey (1971) developed a series of behaviorally based rating scales, which exhibited Guttman scaling properties, for production foremen. More recently, Blanz and Ghiselli (1972) have used Guttman scaling techniques to evaluate the accuracy of supervisors' ratings along 18 performance dimensions. One objective[3] of this research was to develop a set of behaviorally based rating scales, which would exhibit adequate Guttman scaling properties, for systems analysts and programmer/analysts.

A second objective of this research was to investigate the possibility of using Guttman scaling procedures to detect relatively poor raters, that is, supervisors who do not do a particularly good job of rating job performance. It should be recalled that the coefficient of reproducibility in Guttman scaling is a function of the number of rating errors that have been made. An "error" occurs when a response falls outside of the limits of a particular item (or incident) cutting point that has been established during the scaling process and located so as to minimize error (Edwards, 1957). When a large number of errors occur, it is commonly assumed that the incidents comprising the scales were not properly scaled, that is, in the context of a job performance scale, the items did not really reflect a continuum of incidents ranging from high effectiveness to high ineffectiveness. Two other alternatives also exist. First, the items may not be representative of the domain content. The second alternative is that certain raters may be inaccurate. Blanz and Ghiselli (1972) found a wide range in the accuracy of raters using Guttman techniques. Concerning this last alternative, it may be that the rater did not know the ratee well enough to judge his performance. Similarly, the rater may not have been paying sufficiently close attention to the context of the performance items used and thus might have been responding inconsistently. Using a Guttman scaling procedure, it may be possible to examine the errors made over a number of raters and incidents, and eliminate poor incidents and/or poor raters from the performance ratings.

---

[1] Financial support for the analysis of this data was provided by the Center for Business and Economic Research, University of Tennessee.

[2] Requests for reprints should be sent to Richard D. Arvey, Department of Industrial Management, University of Tennessee, Knoxville, Tennessee 37916.

[3] These scales are presently being used as criteria for a synthetic validation research project.

If the tendency to rate poorly generalizes across ratees and across job dimensions, it may be worthwhile to attempt to learn more about the correlates of poor raters. First, however, the stability of the rating characteristic should be investigated by determining whether raters who rate poorly on one dimension also rate poorly on other job dimensions. Further, whether raters who rate one individual poorly also rate other individuals poorly should be determined.

METHOD

Development of Rating Scales

The approach used to develop the scales was essentially a variant of the critical incident methodology as modified by Smith and Kendall (1963). Essentially, the technique involved participation of appropriate personnel in a series of workshops to consider the major performance components or dimensions of the two jobs in question—systems analysts and programmer/analysts. Further, individuals involved in these workshops were asked to generate specific behavioral incidents to serve as anchor points along the rating scales which reflected these dimensions. The general procedure for developing these rating scales has been previously described in some detail (Campbell, Dunnette, Arvey, & Hellervik, in press; Folgi et al., 1971), and the specifics of the development of the present scales are presented elsewhere (Hoyle & Arvey, 1972); therefore, these steps will not be discussed in this study. One point worth noting, however, is the fact that the basic job dimensions of the two job areas were very much the same. Thus, it was decided to develop one set of rating scales to evaluate individuals in these two nominally different jobs. However, the basic task dimensions of the two jobs were suspected to be the same, but what differentiated the two jobs would be the relative importance of the dimensions. This reasoning corresponds closely to the statement by Mayer and Stalnaker (1968).

The results of these series of workshops was the identification of the major dimensions of job behavior of systems analysts and programmer/analysts. These dimensions were as follows: (*a*) technical knowledge, (*b*) planning, organizing, and scheduling, (*c*) maintaining customer relations, (*d*) providing supervision and leadership, (*e*) training others, (*f*) documentation, (*g*) maintaining communications, (*h*) assessing customer needs and providing recommendations, (*i*) job commitment and effort, (*j*) debugging, (*k*) program and/or systems modification and/or development, (*l*) conducting presentations. Two rating scales were developed for each job dimension with the exception of one dimension—conducting presentations—due to the limited behavioral anchors generated for this scale. The first type was simply a rating scale where the behavior incidents were arranged on a continuum. Ratings were made by asking supervisors to infer the level of effectiveness that could be expected of the ratee's typical performance, and to indicate that individual's rating by writing "typical" next to that particular incident.[4] This type of scale was scored by assigning a value of 7 to the top (highly effective) incident down to a value of 1 for the bottom (highly ineffective) incident. Henceforth these scales are referred to as the typical rating scales.

A second type of scale was constructed for each dimension wherein the behavioral incidents were originally arranged in random order on one page, but then reordered to reflect a continuum from high to low effectiveness. Behavioral incidents other than those used in the typical method were used. Raters were instructed to read each incident and decide whether the individual he rated was better than, the same as, or not as good as the individual described in the incidents. The scales were scored by assigning a numerical value to each response category (3 = better than, 2 = the same as, 1 = not as good as) and summing these values over the incidents in each continuum. If the items are properly scaled in a Guttman sense (i.e., no errors), the responses should follow a general pattern—better than rating at the bottom of the scale, change to the same as rating, and finally switch to a not as good as rating somewhere near the top of the scale. Scaling procedures allow for the identification of the incidents and raters where this pattern is not followed. The latter rating technique will be referred to as the Guttman rating method.

---

[4] Raters were also asked to make a rating of "best," indicating the level of performance that could be expected when the rater was at his best. These data were not used in the analysis.

The next phase in the research was to evaluate the psychometric properties of these rating scales.

Sample

Two hundred systems analysts and programmer/analysts (113 programmer/analysts, 87 systems analysts) employed by a large midwestern computer manufacturing company were rated by their supervisors on the above dimensions of job performance using the typical and Guttman rating methods. On each dimension, these supervisors were asked first to make a decision concerning whether the incumbent actually should be evaluated, and if so, to complete his rating on that dimension. If it did not apply, the supervisor left the page blank and went on to the next dimension. Raters were further instructed to indicate an overall effectiveness rating on a 9-point scale. They also rated how important the work dimensions were for the individual job incumbent by allocating 100 points to the different dimensions according to their importance (higher importance = more points). One hundred and three supervisors rated the 200 incumbents.

Procedures

The first set of analyses involved comparing individuals in the two jobs on the importance ratings and the actual ratings made by the supervisors on the dimensions. Thus, means and standard deviations were computed and *t* tests run between the individuals in these two job areas on the importance ratings, the typical, and the Guttman ratings. Internal consistency reliability values for the Guttman rating scales were computed using Cronbach's alpha (Guilford, 1954). It was not possible to estimate the reliabilities of the typical rating scales as the necessary data were not available.

Second, a multidimensional-multimethod correlation matrix was computed in order to examine the convergent and discriminant validity of the rating scales (Campbell & Fiske, 1959). The Guttman scales were subjected to scaling procedures whereby the coefficient of reproducibility for each dimension was computed. The Biomedical Computer Program (BMD 055) was used for this purpose (using the Cornell method) and modified slightly in order to punch out the number of errors made by each rater on each dimension. These error scores were intercorrelated in order to determine whether there was a tendency to rate poorly on more than one dimension. Error scores were also correlated with the measures of job performance. For supervisors who rated two or more individuals, an intraclass correlation coefficient was computed (Winer, 1971) for each dimension to determine the reliability of their error scores across job incumbents. For each supervisor who rated three or more incumbents, a pair of ratees was selected randomly for this last part of the analysis.

RESULTS

Table 1 presents the results of the analysis comparing individuals in the two nominal job areas on the importance ratings of the performance dimensions. There are several aspects to note about these data. First, as was suspected, there were aggregate differences between the two job areas with regard to what is considered important in the jobs. Statistically significant differences emerged on 7 of the 12 dimensions. The dimensions considered more important for programmer/analysts than systems analysts were documentation, debugging, and program and/or systems modification and/or development, while those considered more important for the systems analysts were maintaining customer relations, maintaining communications, assessing customer needs, and conducting presentations.

It appears from these data that, on the average, systems analysts have more to do with personal contact activities, while programmer/analysts deal perhaps at a more technical level. It is important to note, however, the wide variances within each nominal job category with regard to the importance of each job dimension. Thus, for some programmer/analysts, it is important that they maintain customer relations and communications. Similarly, even though statistically significant differences occurred between the two job areas on various dimensions, there is large overlapping between the two distributions as indicated by the overlap percentages (Tilton, 1937) presented in Table 1.

Whereas statistically significant differences appear between programmer/analyts and systems analysts with regard to the rated importance of the job dimensions, there were no differences between the two groups with

TABLE 1

IMPORTANCE RATINGS FOR TWELVE DIMENSIONS OF JOB BEHAVIOR FOR SYSTEMS ANALYSTS AND PROGRAMMER/ANALYSTS

Job behavior dimensions	Ratings						
	Programmer analysts ($N = 112$)		Systems analysts ($N = 87$)		Distribution		
	$\bar{X}$	*SD*	$\bar{X}$	*SD*	d^a	t	Overlap %
Technical knowledge	15.42	9.25	14.52	7.12	.90	.76	97
Planning, organizing, scheduling	9.57	4.91	9.57	4.41	.00	.00	100
Maintaining customer relations	6.56	6.22	10.69	6.05	−4.13	−4.72*	66
Providing supervision and leadership	2.70	6.24	3.34	5.07	−.64	− .80	95
Training others	5.52	4.63	6.61	4.02	−1.09	−1.77	90
Documentation	11.18	5.59	8.45	4.29	2.73	3.90*	80
Maintaining communications	7.05	4.16	8.50	3.72	−1.45	−2.59*	85
Assessing customer needs and providing recommendations	4.30	4.45	5.94	4.40	−1.64	−2.59*	85
Job commitment and effort	11.61	4.74	10.46	4.50	1.15	1.75	90
Debugging	12.13	6.33	8.84	5.90	3.29	3.78*	78
Program and/or systems modification and/or development	10.95	7.31	8.62	7.20	2.33	2.25*	87
Conducting presentations	3.26	2.83	4.62	3.48	−1.26	−2.74*	85

[a] Mean difference.
* $p < .05$.

regard to their rated performance using either of the two rating methods. Therefore, the two groups were combined for further analyses.

The 22 × 22 correlation matrix (2 rating methods, 11 dimensions) comprising the multidimensional-multimethod matrix is presented in Table 2. Also shown are the correlations between the dimensions and the overall effectiveness measure and the internal consistency reliabilities.

In terms of convergent and discriminant validity, Table 2 indicates that adequate convergent validity was achieved. Convergent validity is demonstrated when the correlations between the same dimension using different rating methods are significantly different from zero and reasonably high. The correlations which represent the convergent validity coefficients are circled in Table 2. All of these are significantly greater than zero and thus suggest that both rating methods show agreement or convergence on the 11 dimensions.

Discriminant validity may be defined as the extent to which a dimension or trait is differentiable from other dimensions. Trait measures can be invalidated because they correlate with other measures that are conceptually different and from which they are intended to differ. Discriminant validity is demonstrated by several criteria. First, a diagonal correlation value (validity coefficient) should be higher than the values lying in its corresponding columns and rows in the heterodimension–heteromethod triangles. The logic behind this is that a measure of a dimension should correlate more highly with another measure of the same dimension than with other variables having *neither* dimension nor method in common.

The validity values in Table 2 are higher than in their respective columns and rows (in the heterodimension–heteromethod triangles) on 204 out of 220 comparisons. Dimensions 5 and 8 accounted for 8 out of the 16 occasions where the validity coefficients were equaled or exceeded by these other values.

Another criterion for demonstrating discriminant validity is that a measure of a dimension should correlate higher with an

TABLE 2

CORRELATIONS BETWEEN ELEVEN DIMENSIONS USING TYPICAL AND GUTTMAN RATING METHODS FOR SYSTEMS ANALYSTS AND PROGRAMMER/ANALYSTS

Method	Typical method											Guttman method											
	1	2	3	4	5	6	7	8	9	10	11	1	2	3	4	5	6	7	8	9	10	11	Overall
Typical																							
1																							
2	61																						
3	52	68																					
4	49	70	71																				
5	52	56	55	55																			
6	58	58	56	44	50																		
7	49	60	57	52	44	52																	
8	40	61	54	57	49	49	40																
9	53	51	58	54	53	46	49	48															
10	66	58	53	51	55	57	49	53	57														
11	66	55	49	53	50	55	49	44	54	69													
Guttman																							
1	62	58	51	39	35	49	58	37	46	54	54	(64)											
2	53	76	60	64	52	53	68	52	53	58	60	72	(73)										
3	44	56	57	43	42	43	51	44	48	45	46	63	64	(53)									
4	44	59	44	56	39	41	55	45	52	56	40	53	58	47	(64)								
5	45	59	57	53	47	39	46	47	42	54	53	61	66	61	63	(57)							
6	50	59	48	48	47	57	44	38	36	49	44	53	67	58	56	55	(72)						
7	42	49	42	43	39	36	61	31	35	44	39	51	61	58	46	55	57	(61)					
8	38	49	44	52	29	31	38	48	37	40	40	65	63	49	56	65	47	42	(68)				
9	43	44	51	50	31	36	45	44	61	46	45	66	61	58	57	53	46	49	63	(60)			
10	55	56	47	46	41	48	46	40	46	59	52	70	65	63	57	64	62	53	45	54	(75)		
11	64	60	45	50	50	47	47	45	46	59	65	68	66	57	52	60	53	52	63	52	64	(70)	
Overall	57	71	58	54	52	48	56	49	52	59	54	58	69	50	61	56	46	39	53	45	60	58	

Note. Circled numbers indicate validity diagonal; diagonals with broken lines equal heterodimension-heteromethod triangle, diagonals with solid lines equal heterodimension-monomethod triangle; numbers in parentheses are reliability coefficients using Cronbach's alpha. For typical monomethod triangle, median $R = .55$; for Guttman monomethod triangle, median $R = .61$.

independent effort to measure the same dimension than with measures designed to get at different dimensions that happen to use the same method. For a given dimension, this involves comparing its value in the validity diagonal with its corresponding values in the heterodimension–monomethod triangles. In Table 3, the validity values exceed the values in the corresponding columns and rows in the typical rating method triangle 79 out of 110 comparisons and 59 out of 110 comparisons in the Guttman monomethod triangle. Again, dimensions 5 and 8 do not stand up well under this criterion, exceeding the monomethod coefficients only 8 out of 40 comparisons.

With the exception of these latter two dimensions, we can conclude that the rating scales exhibit reasonably good convergent and discriminant validity. It is interesting to note that the monomethod correlations for the Guttman rating method are generally higher than those in the typical monomethod block. The median intercorrelation between the eleven dimensions using only the Guttman rating method was .61, while the median correlation for the typical method was .55.

The next step in the analysis involved computing the coefficients of reproducibility for the eleven Guttman rating scales. Table 3 presents these coefficients along with the minimum marginal reproducibilities for the eleven dimensions. Guttman (1944) suggests that reproducibilities over .85 indicate that the data have been sufficiently scalable. Guilford (1954) suggests that reproducibilities in the neighborhood of .80 represent quasi scales. Eight of the eleven reproducibilities are over .85, and all of them exceed .80. Another factor to consider when evaluating the scalability of a set of statements is whether the coefficients of reproducibility exceed the minimum marginal reproducibilities. The minimum marginal reproducibility refers to the percentage of nonerror which occurs if the modal response categories of the items were used as the predicted rating. All 11 dimensions exhibit coefficients of reproducibility which exceed their respective minimum marginal reproducibilities, indicating that the rating scales did, for the most part, meet this criterion for scalability.

TABLE 3

COEFFICIENT OF REPRODUCIBILITY AND MARGINAL REPRODUCIBILITIES FOR ELEVEN GUTTMAN RATING SCALES

Guttman rating scales	Reproducibility[a]	
	Coefficient	Minimum marginal
1	.86	.80
2	.82	.74
3	.90	.81
4	.88	.74
5	.83	.75
6	.92	.82
7	.89	.75
8	.88	.80
9	.90	.81
10	.91	.82
11	.84	.72

[a] Coefficient of reproducibility = 1 − # of errors/total number of responses.

The number of rating errors for each rater per dimension was computed. The intercorrelations between the dimensions with regard to the number of rating errors made revealed that 30 out of 55 correlations were significantly different from 0. However, the median correlation was .25 with a range between .41 and −.07. These data inciate that there is a slight tendency for raters who make errors on one dimension to make errors on other dimensions.

It was noted that the number of errors made from a given dimension was related to rated job performance. Visual inspection of the scatterplots did not reveal evidence of nonlinearity. The correlations between the number of errors and the job performance rating (typical rating format) in that dimension were as follows for the eleven dimensions:

−.31, −.24, −.28, −.27, −.33, −.27,
−.11, −.29 −.09, −.31, and −.22.

All but two of these correlations were statistically significant. What these data indicate is that individuals who are rated as doing poorly on a particular aspect of their job are not rated as accurately by their supervisors.

In investigating the reliability of supervisors' error scores when evaluating more than one individual, the resulting intraclass correlation coefficients ranged between −.12 and .45 with a median correlation of .16. The coefficients for dimensions 3, 9, and 10 were significantly different from 0 (McNemar, 1962) but nevertheless quite low (.35, .30, and .26, respectively). In general, the evidence does not point to a strong tendency for raters who make errors on one dimension when evaluating a particular individual to make errors on that dimension when rating another individual.

DISCUSSION

The data presented indicate that it is possible to develop behaviorally based rating scales which exhibit Guttman scaling properties. Further, these scales (along with more traditional behaviorally based rating scales) also exhibit reasonably good convergent and discriminant validities. The finding of convergent and discriminant validity is perhaps more notable when it is recalled that the two rating methods differed not only in their response format, but also in the behavioral incidents that were included in each method. With the exception of two performance dimensions (5 and 8), these scales appear to be appropriate for evaluating performance in these jobs. However, because of the relatively lower intercorrelations among dimensions using the more traditional typical rating method when compared to the Guttman ratings, the typical rating scales might be used as the primary measuring instrument.

The use of these Guttman rating scales to identify relatively poor raters does not seem promising. While there was some tendency for raters who made large numbers of errors on one dimension to make errors on other dimensions, the relationships were weak and practically meaningless. Similarly, there was only slight evidence that raters who made large numbers of errors on a dimension for a particular ratee also made errors when rating another individual on that same dimension. In sum, rating poorly as defined by this technique does not appear to be a generalized trait characteristic of raters. In fact, rater errors are more strongly related to individual ratee job performance in a particular dimension than anything else. Gordon (1970) presented results showing a phenomenon where raters do not note incorrect performance as accurately as they do correct performance, where accuracy referred to the relationship between judges' perceptions of behavior and the actual behavior that took place. What might be occurring is that the poor performers may actually be inconsistent or variable in their job performance. In a slightly different vein, variable performance might also be judged as poor performance by supervisors.

Finally, using ratings to gather information about the importance of each job dimension appears to be a useful technique in understanding similarities and differences among jobs. Significant differences appeared between the two jobs with regard to the rated importance of the various dimensions; however, wide overlap also existed between the two nominally different jobs. It was also noted that the individuals performing in these two jobs differed widely in the patterns of dimensions that were important. Further research should perhaps focus on identifying clusters of individuals who share similar job responsibilities and determining whether these clusters moderate other relationships.

The dimensions which were derived and the ratings of their importance have further implications. Conceivable use could be made of them for job evaluation, wage, and salary purposes. Further, these scales might be used in training and development functions by helping to specify strengths and weaknesses for both the aggregate and individual cases.

REFERENCES

BLANZ, F., & GHISSELLI, E. E. The mixed standard scale: A new rating system. *Personnel Psychology,* 1972, **25**, 185–199.

CAMPBELL, J. P., DUNNETTE, M. D., ARVEY, R. D., & HELLERVIK, L. W. The development and evaluation of behaviorally based rating scales for first level retail managers. *Journal of Applied Psychology,* in press.

CAMPBELL, J. P., DUNNETTE, M. D., LAWLER, E. E., & WEICK, K. E. *Managerial behavior, performance, and effectiveness.* New York: McGraw-Hill, 1970.

CAMPBELL, D. T., & FISKE, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, **56**, 81–105.

DUNNETTE, M. D. *Personnel Selection and Placement*. Belmont, Calif.: Wadsworth, 1966.

EDWARDS, A. L. *Techniques of Attitude Scale Construction*. New York: Appleton-Century-Crofts, 1957.

FOLGI, L., HULIN, C. L., & BLOOD, M. R. Development of first-level behavioral job criteria. *Journal of Applied Psychology*, 1971, **55**, 3–8.

GORDON, M. E. The effect of correctness of the behavior observed on the accuracy of ratings. *Organizational Behavior and Human Performance*, 1970, **5**, 366–377.

GUILFORD, J. P. *Psychometric Methods*. Hightstown, N.J.: McGraw-Hill, 1954.

GUILFORD, J. P. *Fundamental Statistics in Psychology and Education*. Hightstown, N.J.: McGraw-Hill, 1965.

GUTTMAN, L. A basis for scaling qualitation data. *American Sociological Review*, 1944, **9**, 139–150.

HELLERVIK, L. W., DUNNETTE, M. D., & ARVEY, R. D. *Development and pretesting of behaviorally defined job performance measures for foremen in Ford Motor Company Transmission and Chassis and Automotive Assembly Divisions*. (Confidential report submitted to Ford Motor Company) July, 1971.

HOYLE, J. C., & ARVEY, R. D. The development of behaviorally based rating scales for systems analysts, programmer analysts, and computer operators. *Proceedings of the Tenth Annual Computer Personnel Research Conference*, June 15–16, 1972, **10**, 85–103.

MAYER, D. B., & STALNECKER, A. W. Selection and evaluation of computer personnel—the history of SIG/CRP. *Proceedings of the 6th Annual Conference Computer Personnel Research Group*. Johns Hopkins University, 1968, **6**, 657–670.

MCNEMAR, Q. *Psychological statistics*. New York: Wiley, 1962.

SCHULTZ, D. G., & SIGEL, A. I. Generalized Thurstone and Guttman scales for measuring technical skills in job performance. *Journal of Applied Psychology*, 1961, **45**, 137–142.

SMITH, P., & KENDALL, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, 1963, **47**, 149–155.

TILTON, J. W. The measurement of overlapping. *Journal of Educational Psychology*, 1937, **28**, 656–662.

WINER, B. J. *Statistical Principles in Experimental Design*. Highstown, N.J.: McGraw-Hill, 1971.

(Received October 2, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 69–73

DEVELOPMENT OF CRITERION MEASURES OF OVERSEAS MISSIONARY PERFORMANCE[1]

PATTYE WEAVER KENNEDY AND RALPH MASON DREGER[2]

Louisiana State University

To establish criterion measures of effectiveness in overseas missionary performance, a committee of experts selected criterion variables involving personal, social, and professional elements of missionary work and designed The Missionary in Action—A Descriptive Check List (MINA) to measure behavioral attributes relevant to the missionary enterprise. A factor analysis of the MINA check list resulted in 11 factors. The major results of the study indicated that future efforts should be directed toward developing the MINA check list and other criterion measures derived from native cultures.

Church missionary boards have been sending large numbers of missionaries abroad since the beginning of the modern missionary movement in the early part of the 19th century. Millions of dollars are spent annually to recruit, train, and maintain this body of overseas personnel. Despite the involvement of a large number of people and the vast expenditures of money, however, little is known about the quality of services being rendered, the elements which make for effective overseas performance, and the personal characteristics and experiences needed to succeed in missionary service (Cleveland, Mangone, & Adams, 1960). This lack of knowledge regarding the efficacy of the missionary enterprise has prevented the establishment of criteria for measuring overseas missionary performance and, consequently, the selection of predictors of these criteria.

While the investigation of missionary work, specifically, has been the focus of very few research endeavors, extensive efforts have been made to examine other areas of overseas work. A number of sources have suggested the usefulness of psychological examinations (Gunderson & Kapfer, 1966; Hanfmann, 1966; Mischel, 1965), a knowledge of background information (Jones, 1967; Mabry, 1969), and an understanding of situational variables (Mabry, 1969) in reducing the occurrences of incompetence, maladjustment, and breakdown in overseas situations.

However, from a practical standpoint the results of the research efforts have been disappointing. One of the major problems seems to be that investigators have concentrated primarily on evaluating selection procedures without sufficient information pertaining to criteria, so that the results of efforts to establish selection procedures are difficult to interpret and tend to be insignificant (Menninger & English, 1965; Gunderson & Kapfer, 1966). As far as is known, no research has focused specifically on the establishment of criteria for effective overseas missionary performance. From the literature it is also clear that criterion measures should be derived on the basis of information gathered in the native culture rather than in this culture. Ratings made in a home situation and ratings obtained from overseas respondents have been found to vary significantly (Jones, 1967).

For several decades mission boards have recognized the limitations of their selection methods and have felt a need to evaluate their procedures. It was not until 1966, however, that a group of missionary personnel secretaries and psychologists conferred about evaluative procedures. Subsequently, a committee was established within the National Council of Churches of Christ for the purpose of investigating selection procedures (Dreger, 1967). The present work was an outgrowth of the efforts of this committee.

---

[1] The authors wish to acknowledge the support, financial and otherwise, of the National Council of Churches of Christ in America, and to thank the home office missionary personnel secretaries of the various denominations named herein, as well as the missionaries and their colleagues who participated in this study.

[2] Requests for reprints should be sent to Ralph M. Dreger, Department of Psychology, Louisiana State University, Baton Rouge, Louisiana 70803.

The purposes of the total study were to establish criteria for effective overseas missionary service and to examine predictor variables in relation to the criteria. However, this report primarily focuses on the development and use of one instrument, The Missionary in Action—A Descriptive Check List (MINA),[3] for criterion purposes. A report of the total study can be found elsewhere (Kennedy, 1972).

METHOD

Subjects

The subjects were 567 individuals, 137 of whom were missionaries constituting the primary targets and 430 of whom were peers serving as respondents *re* the missionaries. Besides missionaries and peers 10 home office personnel secretaries rated the missionaries, and six served as judges to obtain ideal missionary profiles. The 137 missionaries, men and women, were associated with the United Methodist Church, United Church of Christ of the Philippines, American Baptist Foreign Mission Society, the Episcopal Church, and the Missouri Synod Lutheran Church. The majority of the female missionaries were wives of male missionaries. These subjects were in service in the Philippines or on furlough as of July 1, 1970, or had completed their missionary term, had withdrawn from service, or had been on leave of absence since January 1, 1968.

Description of the MINA Check List

The Missionary in Action check list was designed for this study to measure behavioral attributes specifically relating to personal–social–work relationships of missionaries. Initially, items were derived from suggestions pertaining to qualities of missionary effectiveness made by consultants in the missionary field. These items were supplemented with items representing more general effectiveness of a person in social situations. The total pool of items was subjected to revision, deletion, and/or addition by experts in the missionary enterprise. Where it was felt the terminology would be confusing to the Filipino people the pool of items was further revised in the Philippines on the basis of a translation into Tagalog, a native language. The final revision was given a preliminary tryout in the Philippines prior to this study.

The MINA check list consisted of 155 items on a 5-point scale. A question mark was used in responding if a subject felt he could not rate the missionary on an item.

Administration of Instruments

The MINA check list, along with the FIRO-B (Fundamental Interpersonal Relations Orientation-Behavior), a standardized measure of interpersonal relationships (Schutz, 1967), was administered in the Philippines by the experimenter, the senior author, to the missionaries and peer respondents who had worked closely with the missionaries. Standard instructions were read at the beginning of each session.

Ratings of supervisors on the MINA check list were also obtained for 112 of the missionaries from members of the missionary personnel staff and/or the regional secretaries in the home offices. The remaining 25 missionaries were not well enough known to permit ratings by the home office staff members.

For purposes of comparison, regional and personnel secretaries from six of the various mission boards completed the MINA check list and the FIRO-B in accordance with their conceptions of a hypothetically effective, desirable missionary who might stand within the top 10% to 20% of missionaries. The median of the six profiles on each instrument served as the standard or ideal profile for the MINA and the FIRO-B.

Analyses of the Data

Multivariate procedures were utilized in the major analyses of the data. Since a number of the statistical processes involved results derived from previous analyses, the specific analytic procedures are detailed with the results.

RESULTS

Analysis of the MINA Check List

When respondents had not observed the missionary in situations described by particular items, the response mode used on the MINA check list resulted in missing data. The mean of a respondent's total score was entered in place of the missing information. Then, the scores for the 567 respondents in the Philippines, missionaries and peers, on the 155 items of the MINA check list were intercorrelated, and the resulting matrix was subjected to a principal axes factor-analysis using squared multiple correlation coefficients as diagonal entries (Harman, 1967). Rotation of the derived factors was accomplished through the varimax method, followed by the promax method which yielded an oblique solution.[4]

Although the latent-root-of-one test and Cattell's Scree Test would allow all 30 of the factors the program produced, only 11 of

---

[3] A copy of the Missionary in Action—A Descriptive Check List may be obtained from either author.

[4] Credit for the integration of these procedures should be given to Richard L. Gorsuch of the Kennedy Center, Vanderbilt University.

TABLE 1

DESCRIPTION OF 11 MISSIONARIES IN ACTION CHECK LIST FACTORS AND REPRESENTATIVE ITEMS

Factor	Representative items
A. Describes a person who is very understanding and accepting of people and ideas.	77. People say of him, "He understands me and he sympathizes with me. He knows how to put himself in my position." 106. Christians say they are grateful for his helping them to understand better what it means to love their fellowmen.
B. Describes a person who is insensitive to events around him, has very few, if any, close friends, and has few personal skills for coping with social and professional situations.	78. From the remarks he makes you can tell that he does not see and appreciate the humor in situations. 130. He seems not to be well informed about important local and world developments.
C. Is concerned with an individual's ability to organize his time and energy in carrying out his professional responsibilities.	152. He handles details well without losing sight of the aims of the project in which he is engaged. 24. Whatever he undertakes, he does well.
D. Measures a person's openness to and acceptance of changes in people and social situations.	121. People say that when social changes take place, he reacts positively to them. 84. He helps others develop according to their natural abilities, without forcing them to become what he wants them to be.
E. Deals with the individual's philosophy of life and the way in which this philosophy affects his personal and professional activities.	86. People ask him for advice on both practical and intellectual problems. 142. People appear to enjoy being with him.
F. Is concerned with an individual's leadership abilities.	53. When following rather than leading is called for, he fulfills a follower's role in such a manner as to deserve favorable comment. 31. When leadership is called for, he fulfills a leader's role in such a way as to deserve favorable comment.
G. Consists of items which pertain to an individual's commitment to Christ and his efforts and abilities to share this faith with others.	52. He is spoken of as one who is very deep spiritually. 28. Because of his contact with them, some people who were without Christian faith have become committed to Christ.
H. Taps the traits of humility and dedication.	148. It is clear that he is trying to understand and adjust to life in the Philippines. 140. He gives and takes compliments gracefully.
I. Describes a person who has the capacities to adjust to cultural demands and to relate to people in cultures different from his own.	134. He has close friends among the people of the Philippines. 20. He is generally considered successful in his missionary work.
J. Measures a person's concern about people with special needs, such as the poor, the blind, and the physically handicapped.	21. People feel he has made them more aware of those who have special needs, such as the blind and the physically handicapped. 42. He speaks and acts in such a way as to indicate he is aware of persons with special needs, such as the blind and the physically handicapped.
K. Deals with the family and home relationships of the missionary.	111. (as a married person) There seems to be a calm and pleasant atmosphere among the members of his family. 103. (as a married person) The members of his family manifest their affection for one another.

these could be retained for interpretation. The majority of the remaining 19 factors had factor loadings below .35, the criterion for acceptance, while several were singlets. Table 1 gives a brief description of each of the 11 factors along with representative items.

Reliability coefficients for the MINA check list were computed for the total test and each of the 11 factors from the intercorrelations of items (Nunnally, 1967), according to the procedure rationalized by Dreger (1973) for nonability tests. The results reflected a coefficient of .98 for the total test, with a range of .93–.73 for the factors. Although authorities differ as to whether coefficient alpha and its special cases constitute "lower limits" (Lord & Novick, 1968) or "upper limits" (Nunnally, 1967) to reliability, from an empirical standpoint, it seems that the above coefficients must assuredly represent the latter.

For validational purposes Pearson product-moment correlation coefficients were calculated for each of the 11 factors on the MINA check list between (*a*) median Z scores for each missionary, taken from the scores of the missionary and peer respondents and (*b*) supervisory ratings, derived by having at least three administrators in the home offices designate the top 25% of their missionaries, that is, a rating of 1 was given to each of the missionaries within the top group and a rating of 2 was given to the other missionaries. The coefficients ranged from .289 to .025, with a median of .153. These coefficients served as validity measures along with the r_p coefficients described below.

The factor structure loadings of the 11 factors were used as weights for devising a factor score system for the MINA check list. The resultant 11 factor scores served, thus, as criterion measures for further analyses.

Analysis of MINA and FIRO-B Profiles

For criterion purposes Cattell's (1949) coefficient of pattern similarity (r_p) was calculated for each missionary between his scores and the scores of the peer respondents on the MINA check list and between his scores and the standard MINA profile. Similar r_ps were calculated for the scores on the FIRO-B.

Table 2 presents the mean and median r_p coefficients between self and colleague profiles and between self and standard or ideal profiles on the MINA check list and the FIRO-B. These mean and median coefficients were based on the r_ps of 117 missionaries who completed the MINA check list and 126 who completed the FIRO-B. The results show that the self versus colleague and self versus standard correlations were reversed in magnitude for the MINA check list and the FIRO-B and that the highest relationship existed between self and colleagues on the MINA check list. The comparison yielded a mean of .28 and a median of .40.

Canonical Correlation Analyses

Data from the MINA check list and the FIRO-B were used as criteria along with other criterion measures. The final analyses were computed on 25 criterion variables and 46 predictor variables. Canonical correlation analysis was employed to determine the relation between the predictors and criteria (Cooley & Lohnes, 1962, 1971), and two analyses were performed. However, since the results of these analyses go beyond the scope of this report, they are not given in this study but can be found in Kennedy (1972).

DISCUSSION

One of the major purposes of this study was to establish criterion measures of effective missionary performance. Previous research has pointed to the necessity of deriving criterion measures of effective overseas performance on the basis of information gathered in a foreign culture. The present study attempted to achieve this purpose through extensive data collection in the Philippines.

The effort to develop criterion measures was only partially successful, but a promising aspect of the study relates to the development of the MINA check list. Its construction included the combined efforts of foreign and home office personnel, and it was given an extensive tryout in the Philippines. The factor analysis of the MINA check list produced 11 factors representative of attributes generally considered to be important in social and professional relationships in the missionary situation. The reliability is satisfactorily high. The r_p coefficients between self and colleagues indicated that for these two groups the MINA check list measures similar concepts. The reversal in magnitude of the self versus colleagues and self versus standard r_p correlations on the MINA check list and FIRO-B suggested that the concepts

TABLE 2

SUMMARY OF r_p COEFFICIENTS ON THE MISSIONARY IN ACTION AND THE FUNDAMENTAL INTERPERSONAL RELATIONS ORIENTATION–BEHAVIOR

Category	r_p Coefficient	
	Median	Mean
MINA, self and colleague	.40	.28
MINA, self and standard	.16	.15
FIRO-B, self and colleague	.10	.07
FIRO-B, self and standard	.21	.27

Note. MINA = Missionary in Action—A Descriptive Check List. FIRO-B = Fundamental Interpersonal Relations Orientation-Behavior.

measured by the MINA check list are more relevant to the missionary enterprises and the people who used the MINA check list than the concepts measured by the FIRO-B, an instrument developed for more general purposes. These results provide support for the usefulness of the MINA check list as a tool in measuring certain behavioral attributes and, in turn, the findings support the importance of these behavioral attributes as criterion measures.

It is recognized that the results derived from the use of the MINA check list are tentative and must be replicated in independent studies in order to be interpreted with confidence. It is also recognized that revisions of the instrument are needed. However, the development of any measurement instrument is a long and arduous task. When the process is begun with very little information to build on, as was true in the development of the MINA check list, the difficulties are accentuated. Although the present efforts represent only the beginning phases of the construction process, they do provide information on which to build and provide a rationale for continuing the process.

REFERENCES

CATTELL, R. B. r_p and other coefficients of pattern similarity. *Psychometrika,* 1949, **14**, 279–298.

CLEVELAND, H., MANGONE, G. J., & ADAMS, J. C. *The overseas Americans*. New York: McGraw-Hill, 1960.

COOLEY, W. W., & LOHNES, P. R. *Multivariate procedures for the behavioral sciences*. New York: Wiley, 1962.

COOLEY, W. W., & LOHNES, P. R. *Multivariate data analysis*. New York: Wiley, 1971.

DREGER, R. M. Research design for evaluating the effectiveness of several predictors of acceptance for and success in the Christian missionary enterprise. An unpublished manuscript, Louisiana State University, July, 1967.

DREGER, R. M. A note on the use of absolute values in estimating reliability from the inter-item correlations. *Educational and Psychological Measurement,* 1973, **33**, 3–5.

GUNDERSON, E. K., & KAPFER, E. L. The predictive validity of clinical ratings for an extreme environment. *Brit'sh Journal of Psychiatry,* 1966, **112** (485), 405–412.

HANFMANN, E. Projective techniques in the assessment program of the Office of Strategic Services. In A. Anastasi (Ed.), *Testing problems in perspective*. Washington, D.C.: American Council on Education, 1966.

HARMAN, H. H. *Modern factor analysis*. Illinois: University of Chicago Press, 1967.

JONES, R. R. The predictive validity of background and training assessments in Peace Corps selection. Paper presented at the meeting of the American Psychological Association, Washington, D.C., September, 1967.

KENNEDY, P. W. A pilot study to investigate the procedures used in the selection and evaluation of missionaries. An unpublished dissertation, Louisiana State University, 1972.

LORD, F. M., & NOVICK, M. R. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.

MABRY, H. P. Career change and continuity: Some social and orientative correlates among active and terminated Methodist missionaries, 1946–1967. An unpublished dissertation, Boston University, 1969.

MENNINGER, W. W., & ENGLISH, J. Psychiatric casualties from overseas Peace Corps service. *Bulletin of the Menninger Clin'c,* 1965, **29**(3), 148–158.

MISCHEL, W. Predicting the success of Peace Corps Volunteers in Nigeria. *Journal of Personality and Social Psychology,* 1965, **1**, 510–517.

NUNNALLY, J. C. *Psychometric theory*. New York: McGraw-Hill, 1967.

SCHUTZ, W. C. FIRO-B (*Fundamental interpersonal relations orientation-behavior*). Palo Alto, Calif.: Consulting Psychologists Press, Inc., 1967.

(Received November 13, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 74-78

INTRINSIC VERSUS EXTRINSIC CUES AS DETERMINANTS OF PERCEIVED PRODUCT QUALITY

GEORGE J. SZYBILLO AND JACOB JACOBY[1]

Purdue University

This study tested a portion of Olson's cue model which hypothesizes that, other things equal, intrinsic cues (e.g., physical product differences) will be stronger determinants of perceived-quality judgments than will extrinsic cues (e.g., price or store image). Ninety female undergraduates were randomly assigned to the conditions of a 2 × 3 × 3 (Price × Store Image × Hosiery Product Samples) analysis of variance with repeated measures on the intrinsic factor. Results supported the hypothesis; omega-square values revealed that 73% of the total variance was due to intrinsic product sample differences. Additionally, Shapiro's hypothesis that perceived worth would be more strongly related to stated purchase intentions than would perceived quality was supported.

Purchase behavior in contemporary American society requires, among other things, that the consumer be able to make judgments and comparisons across products, brands, and services. Such judgments are arrived at by a process in which some or all of the various items of information (i.e., cues) associated with the product, brand, or service are identified, evaluated, and integrated to form a composite judgment.

One judgment usually of considerable importance to the consumer is the product's quality, and more than a score of published studies have examined the effects of specific cues on the judgment of perceived quality (cf. Jacoby, Olson, & Haddock, 1971; Olson, 1972). Little thought, however, has been given to developing a conceptual framework for expressing how and when specific cues or types of cues operate to form the composite quality judgment. Until recently, the most notable exception has been Cox (1962), who developed a cue-processing model based upon the notions of predictive and confidence value.

Other investigators (Jacoby et al., 1971; Valenzi & Andrews, 1971) have distinguished between intrinsic and extrinsic cues. The former refer to cues which, if changed, would produce a resultant change in the physical product itself (e.g., flavor and aroma cues for beer), while the latter refer to such factors as price and store image. While neither Valenzi and Andrews (1971) nor Jacoby et al. (1971) have tested any a priori hypotheses regarding cue-intrinsicness–extrinsicness, the latter speculate, on a post hoc basis, that intrinsic cues might be more important determiners of perceived quality than extrinsic cues.

The most explicit and detailed conceptualization of the quality cue selection and integration process has been developed by Olson (1972; Olson & Jacoby, 1972). One of the numerous propositions contained in this model is that, other things equal, intrinsic cues will be more strongly related to quality perception than will extrinsic cues. The primary purpose of the present investigation was to provide an a priori test of this hypothesis using a product possessing different cues than were involved in the earlier studies. The intrinsic cues employed by Jacoby et al. (1971) and Valenzi and Andrews (1971) were primarily olfactory and gustatory in nature, while the intrinsic cues for the product used in this study (ladies nylon hosiery) were primarily tactile.

Price and store-image information were the extrinsic cues utilized in the present investigation. Only two prior studies examined the effects of price and store image in combination as cues to product quality (Andrews & Valenzi, 1971; Stafford & Enis, 1969). However, subjects in both studies were exposed to all experimental treatment conditions, thereby suggesting that biasing-demand characteristics (Orne, 1962) may have been introduced. In contrast, repeated measures were taken over only one treatment factor in the present study. Moreover, neither of these studies permitted actual product differences to vary as a third factor in the experimental design. Data from Jacoby et al. (1971) suggest that such procedures may oversimplify the quality perception process.

---

[1] Requests for reprints should be sent to Jacob Jacoby, Department of Psychological Sciences, Purdue University, Lafayette, Indiana 47907.

Finally, Shapiro (1970) presented evidence to indicate that perceived worth (i.e., value for the money) has a stronger relationship to perceived likelihood of purchase than does perceived quality. Inasmuch as unbranded product samples were used in the current investigation, the relationships between perceived quality, perceived worth, and likelihood of purchase were also examined. The specific hypotheses of the study were as follows:

Hypothesis 1. Intrinsic cues (i.e., in the present study, composition differences in three different samples of hosiery) exert a greater effect on quality perception judgments than extrinsic quality cues (i.e., in the present study, price and store image).

Hypothesis 2. Perceived worth, defined as value for the money, is more strongly related to stated likelihood of purchase than is perceived quality.

METHOD

Subjects

Subjects were 90 female Purdue University undergraduates enrolled in an introductory psychology course during the fall of 1971. They were required to have lived in the Lafayette area for a minimum of six months in order to insure that the independent variable of store image was meaningful for them.

Design

A 2 × 3 × 3 factorial consisting of two levels of price (i.e., present vs. absent), three levels of store image (i.e., high-quality store image, low-quality image, and no information provided as to store image), and three levels of product samples (i.e., three physically different samples—different brands—of hosiery) was employed. Repeated measures were taken across the latter factor (Winer, 1971, p. 559). Fifteen subjects were used in each of the six between-subject conditions.

Independent Variables

Price information was manipulated by either presenting or withholding the actual prices ($0.98, $1.67, and $2.49) of the three product samples.

Store-image information was manipulated by either presenting or withholding the names of several local stores purportedly selling the three product samples. Two stores were selected on the basis of a pretest which indicated that they represented high- and low-quality store images. The difference between the store with the highest pretest mean quality rating ($\bar{X}_H = 7.95$, on a 9-point scale) and the store with the lowest mean quality rating ($\bar{X}_L = 2.56$) was highly significant ($t = 16.58$, $df = 22$, $p < .001$).

Each product sample of hosiery was identified by either I, J, or R, since a pretest showed no differences in preference across these three letters. The letters were randomly assigned to product samples (i.e., hose) for each experimental condition. For purposes of a simplified exposition, the letters I, J, and R will hereafter be used to designate the low-, medium-, and high-quality product samples, respectively.

The three samples of hosiery were pretested on the attributes of knit, texture, workmanship, cut, and overall quality. Results of an analysis of variance of overall perceived-quality ratings indicated a significant main effect for samples ($F = 24.97$, $df = 2/26$, $p < .0001$), and Newman-Keuls analysis indicated that all differences between pairs of means for the three samples ($\bar{X}_I = 34.28$, $\bar{X}_J = 54.64$, $\bar{X}_R = 84.50$) were significant beyond the .01 level. These differences were in the anticipated direction.

Dependent Variables

The major dependent variable for this study was overall perceived quality. The minor dependent variables were perceived worth and stated likelihood of purchase. All were measured using the 100-point scale procedure described in Jacoby et al. (1971).

Procedure

Subjects in each of the six cells were run separately. The subjects were seated at a long table and separated by table-top partitions to restrict between-subject communication. A female introduced the experiment as a psychological study designed to develop reliable instruments for measuring hosiery perceptions. She then gave the subjects a questionnaire booklet and demonstrated how to use the rating scales. Subjects were told they would be evaluating three samples of hosiery available locally. Hosiery samples were presented in wrapped gift boxes. Price information and store-image information were given, as appropriate, by labels attached to the samples and communicated verbally before subjects began evaluating the hose. Samples were checked after each experimental session for damage, replaced if necessary, and new labels assigned.

RESULTS

Manipulation Check

The manipulation check for store image indicated a significant main effect for quality of store merchandise ($F = 53.02$, $df = 2/87$, $p < .001$). A Newman-Keuls analysis indicated that all differences between pairs of means for the three store-image conditions ($\bar{X}_L = 3.80$, $\bar{X}_O = 5.73$, $\bar{X}_H = 7.93$) were significant beyond the .01 level. These differences were in the anticipated direction, with the store names selected on the basis of the pretest to represent low store image and high store image receiving the lowest and highest mean ratings, respectively.

Model for Data Analysis of Dependent Measure

The model for repeated measures analysis of variance is appropriate if the following assumptions are satisfied: (*a*) homogeneity of the variance–covariance matrices and (*b*) symmetry of the pooled variance–covariance matrix (Winer, 1971, pp. 561–562). For the dependent measure, tests for homogeneity of variance for variation due to subjects within groups and due to the interaction between samples and subjects within groups were computed. The F max tests for homogeneity of variance yielded insignificant results for variation due to subjects within groups ($p > .05$) and due to interaction between samples and subjects within groups ($p > .05$). However, since complete assumptions about the variance–covariance matrices were not tested, the Greenhouse and Geisser (1959) procedure was used for all tests involving the samples factor. This procedure is negatively biased in that it uses conservative degrees of freedom, making rejection of false null hypotheses less likely.

TABLE 1

MEANS AND STANDARD DEVIATIONS OF OVERALL PERCEIVED-QUALITY RATINGS

Quality rating	Sample					
	Price present			Price absent		
	I	J	R	I	J	R
Low image						
$\bar{X}$	24.80	36.46	85.13	19.80	49.86	82.46
SD	20.35	21.64	11.33	13.34	12.87	14.21
No image						
$\bar{X}$	26.46	54.73	82.73	24.73	51.33	86.86
SD	15.25	13.04	15.60	12.43	14.45	7.42
High image						
$\bar{X}$	30.00	57.33	87.06	34.40	56.86	86.66
SD	18.51	15.33	9.98	23.74	17.38	13.37

Note. $n = 15$ per cell.

TABLE 2

ANALYSIS OF VARIANCE OF OVERALL PERCEIVED-QUALITY RATINGS

Source	*df*	*MS*	*F*
Between subjects			
Price (P)	1	56.95	<1
Store image (SI)	2	1810.74	5.47*
P × SI	2	29.47	<1
Error between groups	84	331.13	
Within subjects			
Product samples (PS)	2 (1)[a]	77582.89	394.86**
P × PS	2 (1)	93.36	<1
SI × PS	4 (2)	282.54	1.44
P × SI × PS	4 (2)	417.58	2.13
Error within groups	168 (84)	196.48	

[a] Conservative *df* used in Greenhouse and Geisser procedure.
* $p < .01$.
** $p < .001$.

Data Analysis of Major Dependent Variable

The means and standard deviations of overall perceived-quality ratings for each experimental cell are presented in Table 1. Results of the analysis of variance based on these data are summarized in Table 2. A strong product samples main effect was obtained ($F = 394.85$, $df = 1/84$, $p < .0001$), and a Newman-Keuls analysis indicated that all differences between the pairs of means for the three samples ($\bar{X}_I = 26.76$, $\bar{X}_J = 51.10$, $\bar{X}_R = 85.15$) were significant beyond the .01 level. The analysis of variance also indicated a significant store image main effect ($F = 5.46$, $df = 2/84$, $p < .01$). A Newman-Keuls analysis indicated that only the difference between the means for the low-store-image ($\bar{X}_L = 49.75$) and high-store-image ($\bar{X}_H = 58.72$) conditions was significant ($p < .05$).

Intrinsic quality cues were predicted to exert a greater effect on quality perception than extrinsic cues. To test this hypothesis, omega-square values were computed for the effects of price, store image, and product samples (Winer, 1971, p. 428). In terms of these indices, 73% of the total variance was due to the main effect of samples, and approximately 1% was due to the main effect of store image.

Price information was relatively ineffective in explaining any percentage of the total variance. This finding confirms Hypothesis 1.

The price-present, store-image-absent condition of this study was analogous to Shapiro's (1970) experimental setting. The correlations between perceived quality and stated likelihood of purchase for samples I, J, and R were .07, −.03, and .49, respectively. The correlations between perceived worth and stated likelihood of purchase for these samples were .72, .56, and .64, respectively. Based on tests for differences between dependent correlations (Bruning & Kintz, 1968, p. 193), perceived likelihood of purchase was more closely related to perceived worth than was perceived quality for samples I and J ($t_I = 2.68$, $t_J = 2.60$, $df = 12$, $p < .05$) but not for sample R ($t_R = .75$). These results partially confirm Hypothesis 2.

DISCUSSION

Consistent with earlier research (Valenzi & Andrews, 1971; Jacoby et al., 1971), intrinsic cues had a greater effect upon quality perception than did extrinsic cues. More specifically, composition differences in hosiery samples had a greater effect (accounting for 73% of the quality rating variance) than did either price or store-image information. In contrast to previous studies, the amount of variance accounted for by product composition differences in this study is extremely large. This may be an effect of moving from products for which between-brand differences, although discriminable, have a relatively small range (i.e., butter and beer) to a product for which brand differences appear to have a relatively large range (i.e., hosiery).

Comparing the relative strength of price and store image, the results indicate that while the effect of price was not significant, the awareness of store image did influence quality perception. Apparently, knowing the samples came from a store which carries high-quality merchandise increases quality ratings relative to knowing that the samples came from a store which carries low-quality merchandise. Although the store image effect was statistically significant, its practical importance is mitigated by the small amount of variance it explained. Once again, price was shown to have an insignificant effect on judgments of perceived quality when it was one of several cues in a multicue setting (cf. Jacoby et al., 1971).

Shapiro's (1970) hypothesis that perceived worth is more strongly related to stated likelihood of purchase than is perceived quality received partial support, and this has substantial implications for practitioners. More specifically, and given that stated purchase intentions have some positive relationship to actual purchase behavior, these results suggest that products should not be designed to be of highest quality as much as they should strive to provide the best (perceived) value for the money.

This investigation represents but one small test of Olson's (1972) quality perception model. Additional constructs and hypotheses remain to be examined. Hopefully, this model and investigations such as this will serve to stimulate the research and additional conceptualization necessary for arriving at a comprehensive understanding of the quality perception process.

REFERENCES

ANDREWS, I. R., & VALENZI, E. R. Combining price, brand name, and store cues to form an impression of product quality. *Proceedings of the 79th Annual Convention of the American Psychological Association*, 1971, **6**, 649–650. (Summary)

BRUNING, J. L., & KINTZ, B. L. *Computational handbook of statistics*. Glenview, Ill.: Scott, Foresman, 1968.

COX, D. F. The measurement of information value: A study in consumer decision-making. In W. S. Decker (Ed.), *Emerging concepts in marketing*. Chicago: American Marketing Association, 1962.

GREENHOUSE, S. W., & GEISSER, S. On methods in the analysis of profile data. *Psychometrika*, 1959, **24**, 95–112.

JACOBY, J., OLSON, J. C., & HADDOCK, R. A. Price, brand name, and product composition characteristics as determinants of perceived quality. *Journal of Applied Psychology*, 1971, **55**, 570–579.

OLSON, J. C. Cue utilization in the quality perception process: A cognitive model and an empirical test. Unpublished doctoral dissertation, Purdue University, 1972.

OLSON, J. C., & JACOBY, J. Cue utilization in the quality perception process. *Proceedings of the 2nd Annual Convention of the Association for Consumer Research*, 1972, **2**, 167–179.

ORNE, M. T. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist,* 1962, **17**, 776–783.

SHAPIRO, B. P. The effect of price on purchase behavior. In D. L. Sparks (Ed.), *AMA Fall Educators Conference.* Chicago: American Marketing Association, 1970.

STAFFORD, J. E., & ENIS, B. M. The price–quality relationship: An extension. *Journal of Marketing Research,* 1969, **6**, 456–458.

VALENZI, E. R., & ANDREWS, I. R. Effects of price information on product quality ratings. *Journal of Applied Psychology,* 1971, **55**, 87–91.

WINER, B. J. *Statistical principles in experimental design.* New York: McGraw-Hill, 1971.

(Received September 5, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 79-83

RELATIONSHIP OF INFORMATION-PROCESSING ATTITUDE STRUCTURES TO PRIVATE BRAND PURCHASING BEHAVIOR[1]

JAMES R. BETTMAN[2]

Graduate School of Management, University of California, Los Angeles

Variables related to consumer information-processing models and consumer attitude structures were used to discriminate between private and nonprivate brand purchasers. Discriminant analysis was used to derive discriminant functions and to classify subjects. In general, variables reflecting lower risk and greater information were associated with private brand choices. Approximately 80% of the subjects in the three product classes studied were correctly classified. A scrambled sample validation procedure showed that these results were not spurious. Arguments for the general usefulness of an information-processing and decision-structure-oriented approach were advanced.

There have been several research studies that have attempted to characterize those consumers who purchase private brands and to distinguish them from those who buy nonprivate brands. To date, the results have not proven very successful. For example, Frank and Boyd (1965), defining private brands as those owned by either retailers or distributors and nonprivate brands as those owned by manufacturers, find virtually no predictive power for demographic and socioeconomic variables in attempting to predict the proportion of a household's private brand purchases. They conclude that private brand and nonprivate brand consumers are essentially indistinguishable with respect to such demographic measures. Similarly, Myers (1967), in attempting to predict private brand attitude, found very little predictive power for general personality variables and socioeconomic factors. Finally, Burger and Schott (1972) again found that demographic variables did not discriminate private brand from nonprivate brand buyers. However, they did find some degree of discrimination with attitudinal variables.

These results stem perhaps from the lack of a proper analytic framework. It is certainly easier to form market segments based on demographic and socioeconomic variables, but such variables have not added much to the understanding of the psychology of consumer choice processes. There is no reason to expect demographic or general personality trait measures to correlate well with behavior, as has been shown in many consumer studies (Kassarjian, 1971). Fishbein (1967) has pointed out the same type of reasoning in studying attitude–behavior linkages. General measures of attitude may be compatible with many different behavioral responses; hence, such a general measure should not be expected to correlate highly with any given response.

In contrast, the more specific information-processing models have proven successful in modeling buyer choices in several areas (Bettman, 1970; Haines, 1969). Such models utilize complex decision networks composed of product perceptions and situational variables. Haines (1969) has noted the similarities between this approach and the notion of complex attitude structures developed by Schroder, Driver, and Streufert (1967). They see attitude as consisting of two types of properties:

> An attitude is defined in terms of content properties (magnitude and direction; for example, x degree of conservatism) *and* structural properties (the complexity of conceptual rules involved in processing information about stimuli relevant to the attitude area) [p. 129].

These structural properties of attitude seem to be what is depicted in the consumer decision net information-processing models. As

---

[1] This research was partially supported by the Division of Research of the Graduate School of Management, University of California, Los Angeles.

[2] Requests for reprints should be sent to James R. Bettman, Graduate School of Management, University of California, Los Angeles, California 90024.

Schroder et al. (1967) state, "'structural variables' provide a metric for measuring the way a person combines information . . . for adaptive purposes [p. 4]." Thus, consumer attitude structure properties represent the internal model used by the consumer to process data about stimuli relevant to product choice.

This suggests that, in general, developing an information-processing model of a choice process can conceivably be considered as similar to the problem of assessing attitude–behavior relationships, when attitude is represented by both content and specific attitude structural properties. Fishbein (1972) has found that a fruitful approach to studying attitude–behavior relations involves using situation-specific attitudes and measures of attitude–belief components. This suggests an approach to the problem of discriminating between private and nonprivate brand purchasers. Useful variables for such differentiation are likely to be both related to the content properties of attitude and contained in the situation-specific attitude structures of purchasing. Bettman (1970) found that perceived risk of a product class was an important part of consumer information-processing structures for grocery products. Perceived product quality, informational variables, and the influence of children were also found to be relevant in these models. Note that such variables also should affect the content of the attitude toward grocery product brands, not merely be a part of the structural properties of the attitude that affects consumer information handling.

The objective of this study is to reexamine the phenomenon of private brand purchasing from a different perspective than that of explaining behavior by using general traits and descriptors. In particular, a situation-specific information-processing approach is taken. This study tests the usefulness of this information-processing viewpoint. Those variables previously found to be contained in the attitude structures active in processing information in the grocery brand purchase situation were used to attempt to discriminate between private and nonprivate brand choices. To the extent that this more micro and situation-specific approach is successful, the information-processing approach is supported.

METHOD

Subjects

The subjects for the present study were 123 housewives selected from the Los Angeles area. The subjects were mostly wives of students living in the married student housing complex at the University of California, Los Angeles and members of a women's group. The subjects completed a paper-and-pencil instrument measuring the information-processing variables and some behavioral variables for each of nine types of grocery products. Specifically, the type of brand purchased and the attitude structure variables were obtained as outlined below.

Variables

The subjects were asked to check which of a list of brands in a product class they usually purchased. This allowed determination of the dependent variable of private brand or nonprivate brand choice. There were large enough groups of private brand purchasers to analyze for three of the nine product classes: paper towels, aspirin, and margarine.

The 10 independent variables measured fell into four main types of variables which had previously been found in consumer information-processing attitude structures for grocery products: demographic, perceived risk related, perceived product quality related, and information related (Bettman, 1970). Only 1 demographic variable was used, number of children living at home.

For the perceived risk variables, a measure of the inherent riskiness in a product class was developed by using paired comparisons with a 10-point intensity scale. For each comparison, each subject rated which of a pair of product classes would be more risky to purchase in an imaginary store where all brand levels and prices were disguised. The subject then rated how much more risky the chosen class would be on the intensity scale (with 0 meaning equally risky and 9 meaning much more risky). These data for each subject were then processed as follows to obtain a risk scale. Suppose product classes i and j were the two classes in the pair, and denote the numerical rating obtained by d_{ij}. The following sign convention was used: if i is more risky than j, $d_{ij} > 0$; if i and j are equally risky, $d_{ij} = 0$; and if i is less risky than j, $d_{ij} < 0$. Since this rating was obtained for all the distinct product class pairs only (i.e., $9 \cdot 8/2 = 36$ pairs), set $d_{ji} = -d_{ij}$. Then scale values for each product class s_i were obtained by setting $s_i = \Sigma j \neq i \; d_{ij}/9$.[3] Thus, the s_i represent the relative degree of risk for that subject for each product class; that is, the s_i for each subject summed to zero over the nine classes. In addition, the measures of certainty and danger components of risk used by Cunningham (1967) were used, adapted to 10-point rating scale format.

---

[3] This is akin to Thurstonian scale values. For an example of the use of this technique, see Bechtel, Tucker, and Chang (1971, pp. 374–375).

To measure variables related to perceived product quality, the subjects were given a set of brand names for each product class (each subject had the same set, but the number of brands given differed across product classes). The subjects then placed these brands on a 20-point quality scale, ranging from very low to very high quality. The subjects also marked the lowest level of quality for each product class that would be acceptable to them. Given this data, the following measures were developed for each product class. Mean quality was measured as the mean for the quality levels marked for the sample of brands. Relative variance was obtained by measuring the sample variance in quality for each class and dividing by the maximum variance in quality for each subject over all nine product classes. The acceptance level was measured simply as the quality level marked as the minimal acceptable level. Finally, percent acceptable was defined as the percentage of the sample of brands for each class rated equal to or above the minimal acceptable level of quality by each subject.

Two informational variables were used. Mean familiarity was obtained by having subjects rate each brand in a product class on a 10-point familiarity scale. The average rating across brands was then taken. Confidence in information was measured on a 10-point rating scale for each product class.

Method of Analysis and Hypotheses

Because the dependent variable of private brand purchasing was a binary variable, a two-group multiple discriminant analysis was performed using the 10 independent variables defined above. This analysis yielded both discriminant functions and a classification table for each of the three product classes analyzed.

As discussed above, the information-processing variables measured should be related to consumer choice behavior, since these variables are contained in the attitude structures influencing consumer choice, and should also be related to the content properties of the attitudes invoked in the choice situation. In particular, private brand purchasing should be associated with a lower degree of risk and uncertainty attributed to a product class, higher and less variable perceived quality for brands in the class, and more information and familiarity with the product class. These hypotheses follow from the general notion that private brand purchasing is more likely when the consumer is confident she can obtain satisfactory performance as well as save money. Also, the greater the number of children at home, the greater the familiarity with the class and the motivation to economize on purchases are likely to be. Frank and Boyd (1965) found a slight association between family size and private brand purchases. These general arguments may be translated into specific terms by hypothesizing expected signs for the variables in the discriminant functions. Private-brand-purchasing group membership should be associated with lower inherent riskiness (−); greater certainty (+); less perceived danger (−); higher mean quality and percent acceptable (+); lower acceptable level and relative variance (−); greater mean familiarity and confidence (+); and more children at home (+).

RESULTS

The discriminant functions are presented in Table 1 and the classification results in Table 2. The effective sample sizes for each product class were smaller than 123 because some subjects did not purchase in a product class or purchased brands not listed in the sample of brands used in the study.

Generalized Mahalanobis distance was used to test the significance of the discriminant functions.[4] All three functions were significant at $p < .001$. Note that, in general, the variable coefficients are as hypothesized. One must be careful here, however, since many of the coefficients are not significantly different from zero. Finally, several of the coefficients have what appears to be the wrong sign because they are acting as suppressor variables (Darlington, 1968); that is, the variable actually correlates in one direction with the dependent variable but has the opposite sign in the equation because it is correcting for error in another variable. This is true for percent acceptable for paper towels, acceptable level for aspirin and margarine, mean quality for margarine, mean familiarity for margarine, and confidence for paper towels. The main counter-hypothesis finding seems to be for mean quality. By examining the magnitude of the standardized coefficients, it can be seen that, in general, certainty and percent acceptable are most important for discrimination, although the results vary over product class.

The fact that the discriminant functions are statistically significant does not mean that group members will be correctly classified. In the present study, however, the functions classify quite well, as can be seen from Table

---

[4] Generalized Mahalanobis distance can be used to test the hypothesis that the mean vectors for the two groups are equal. Under this null hypothesis, the distribution of this quantity approaches that of chi-square. For more details, see Afifi and Azen (1972, especially pp. 235–250).

TABLE 1

PRIVATE VERSUS NONPRIVATE BRAND DISCRIMINANT FUNCTIONS

Variable	Coefficients for the product classes		
	Paper towels ($n = 115$)	Aspirin ($n = 113$)	Margarine ($n = 118$)
Constant	.957	.464	−.113
Children at home	.059 (.148)*	.013 (.030)	.004 (.010)
Inherent risk	−.002 (−.007)	−.008 (−.064)	−.024 (−.159)
Certainty	.056 (.355)***	.025 (.162)	.030 (.196)*
Danger	−.008 (−.043)	−.034 (−.221)*	−.033 (−.193)*
% acceptable	−.177^a (−.111)	.388 (.260)*	.475 (.377)**
Relative variance	−.178 (−.115)	−.273 (−.189)**	−.094 (−.073)
Acceptance level	−.032 (−.220)	.007^a (.061)	.035^a (.310)*
Mean quality	−.029 (−.099)	−.050 (−.247)**	−.034^a (−.174)
Mean familiarity	−.055 (−.229)**	.018 (.057)	−.008^a (−.034)
Confidence	.020^a (.116)	.018 (.114)	.021 (.147)
Mahalanobis D^2	39.991***	90.363***	45.682***

Note. Standarized coefficients are given in parentheses.
[a] Variable acting as a suppressor variable.
* $p < .10$.
** $p < .05$.
*** $p < .01$.

2. The percentages correctly classified are 77.4% for paper towels, 82.3% for aspirin, and 80.5% for margarine. However, extreme care must be used in interpreting these results.

First, since the groups of private and nonprivate brand purchasers are of unequal size, the appropriate chance classification model must be determined as a standard of comparison. Morrison (1969) argues that the appropriate model is normally the proportional chance model. If p is the true proportion of private brand purchasers in the sample and α is the proportion actually classified as private brand purchasers, then the probability of a correct classification by chance and, hence, the chance proportion correct is given by

$$P(\text{correct}) = p\ \alpha + (1 - p)(1 - \alpha).$$

TABLE 2

CLASSIFICATION OF SUBJECTS USING THE DISCRIMINANT FUNCTIONS

Classification	Product class		
	Paper towels	Aspirin	Margarine
Actual private brand (classed private brand)	23	23	15
Actual private brand (classed nonprivate brand)	5	7	7
Actual nonprivate brand (classed nonprivate brand)	66	70	80
Actual nonprivate brand (classed private brand)	21	13	16
% correctly classified	77.4	82.3	80.5
% private brand correctly classified	82.1	76.7	68.2

For the three product classes of this study, the proportional chance figure is 56.0% for paper towels, 58.5% for aspirin, and 64.9% for margarine. Thus, the results were substantially better than this chance criterion.

The discriminant analysis of this study used all the data to estimate the discriminant coefficients because of interest in obtaining the most stable estimates of the coefficients possible, given the relatively small sample. As Frank, Massy, and Morrison (1965) point out, if the function based on all the data is then used to classify these same data, there is an upward bias in the percentage correctly classified. Since splitting the sample was undesirable for coefficient estimation purposes, the modified V_2 validation procedure suggested by Frank et al. (1965, p. 255) was applied. The raw data were "scrambled" by assigning individuals to the private brand and nonprivate brand groups at random, while keeping the sizes of the two groups equal to their actual values. Discriminant analysis was then performed on this sample. This procedure was carried out 10 times for each product class. The average percentage correctly classified for each class was 63.5% for paper towels, 61.9% for aspirin, and 63.5% for margarine. Since the actual results are all substantially higher, the results do not seem spurious because of bias.

Finally, Morrison (1969, pp. 160–161) points out that the smaller group is usually difficult to classify well, although it is often the group of interest. In that light, the results for the private brand group are particularly impressive, since 82.1%, 76.7%, and 68.2% are classified correctly for paper towels, aspirin, and margarine, respectively.

DISCUSSION

The results bear out the contention that the perspective of information processing, with its emphasis on the attitude structural properties used to process information by the consumer and the variables contained in these structures, is more suited to analyzing consumer behavior than approaches based on ad hoc demographic or general personality trait approaches. That is, perceived risk, informational, and perceived-product-quality variables are seen as parts of the information-processing attitude structural properties that in conjunction with attitude content influence behavior. The ability of the discriminant analysis to differentiate among the private and nonprivate brand groups bears out this line of argument.

In general psychological terms, this type of viewpoint argues for closer attention to models of choice situation perceptions and choice processes. In this sense, the perspective is similar to that of Fishbein (1967, 1972) in arguing that relationships with behavior will be found only if the actual information-processing aspects of the choice situation are investigated and variables related to the active attitude structures in the situation are considered.

REFERENCES

AFIFI, A. A., & AZEN, S. P. *Statistical analysis: A computer oriented approach.* New York: Academic Press, 1972.

BECHTEL, G. G., TUCKER, L. R., & CHANG, W. C. A scalar product model for the multidimensional scaling of choice. *Psychometrika,* 1971, **36**, 369–388.

BETTMAN, J. R. Information processing models of consumer behavior. *Journal of Marketing Research,* 1970, **7**, 370–376.

BURGER, P. C., & SCHOTT, B. Can private brand buyers be identified? *Journal of Marketing Research,* 1972, **9**, 219–222.

CUNNINGHAM, S. M. The major dimensions of perceived risk. In D. F. Cox (Ed.), *Risk taking and information handling in consumer behavior.* Boston: Harvard Business School, 1967.

DARLINGTON, R. B. Multiple regression in psychological research and practice. *Psychological Bulletin,* 1968, **69**, 161–182.

FISHBEIN, M. Attitude and the prediction of behavior. In M. Fishbein, (Ed.), *Readings in attitude theory and measurement.* New York: Wiley, 1967.

FISHBEIN, M. The search for attitudinal–behavioral consistency. In J. B. Cohen (Ed.), *Behavioral science foundations of consumer behavior.* New York: Free Press, 1972.

FRANK, R. E., & BOYD, H. W. Are private-brand-prone grocery customers really different? *Journal of Advertising Research,* 1965, **5** (4) 27–35.

FRANK, R. E., MASSY, W. F., & MORRISON, D. G. Bias in multiple discriminant analysis. *Journal of Marketing Research,* 1965, **2**, 250–258.

HAINES, G. H. Information and consumer behavior. (Working Paper 6908) Rochester, N.Y.: University of Rochester, College of Business Administration, July 1969.

KASSARJIAN, H. H. Personality and consumer behavior: A review. *Journal of Marketing Research,* 1971, **8**, 409–418.

MORRISON, D. G. On the interpretation of discriminant analysis. *Journal of Marketing Research,* 1969, **6**, 156–163.

MYERS, J. G. Determinants of private brand attitude. *Journal of Marketing Research,* 1967, **4**, 73–81.

SCHRODER, H. M., DRIVER, M. J., & STREUFERT, S. *Human information processing.* New York: Holt, Rinehart & Winston, 1967.

(Received July 31, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 84–86

SHORT NOTES

PREDICTION OF ARTISTIC PERFORMANCE FROM BIOGRAPHICAL DATA[1]

LAWRENCE R. JAMES,[2] ROBERT L. ELLISON, DAVID G. FOX
Institute for Behavioral Research in Creativity, Salt Lake City, Utah
CALVIN W. TAYLOR
University of Utah

A biographical inventory was constructed and administered to 312 nonart students and 501 art students in the areas of music, dance, theater, and the visual arts. The biographical inventory had significant average cross-validities against an art versus nonart criterion (.67, $p < .01$) and multiple art criteria (art sample only) both within and across artistic areas. A tentative conclusion was that prediction of art versus nonart potential and prediction of actual performance in the arts require different scoring procedures for the biographical inventory.

Relatively little research has been conducted on the prediction of different criteria of performance in the arts (Anastasi, 1961). For those studies that have been conducted, however, results have been promising (Anastasi & Schaefer, 1969; Burkhart, 1962). Schaefer and Anastasi (1968), for example, demonstrated that biographical information could predict a "creative artistic" versus a "control artistic" dichotomous criterion for adolescent boys. No studies were found on the prediction of several different criteria common to different art areas.

The present study examines the concurrent cross-validities of a biographical inventory developed to predict an art versus nonart dichotomous criterion as well as several different performance criteria for art students in four art areas (music, dance, theater, and the visual arts). It was assumed that two different biographical inventory scoring procedures would be required, one for differentiating between art and nonart potential and one for predicting actual performance in the arts. This assumption was tested by comparing the cross-validities against art performance criteria for both of the above scoring procedures.

METHOD

Samples

The biographical inventory was administered to 835 art students and 363 nonart students. Return rates were 60% for the art sample and 86% for the nonart sample. The art sample ($n = 501$) obtained was comprised of 106 subjects from a private high school specializing in the arts, 160 subjects from a state summer school for exceptional art (or academic) high school students, and 130 and 105 subjects, respectively, from two urban high schools offering programs in the arts. The nonart sample ($n = 312$) was comprised of 33 subjects from the state summer school and 107 and 182 subjects, respectively, from two urban high schools.

No attempt was made to match the art and nonart samples on such biographical data as experience in the arts and sex, because the investigators were concerned with examining existing differences. However, the only practical difference between the two samples was a significantly higher proportion of females in the art sample (.61. vs. .41).

Biographical Data

By using either the original wording or wording revisions for art students of items developed in an earlier study concerned with predicting scientific and engineering performance (Taylor & Ellison, 1967), 249 biographical inventory items were made available for this study. The biographical inventory items were separated into the following two categories: (*a*) 200 general biographical items (e.g., developmental history, parents and family life, extracurricular activities, etc.) and (*b*) biographical items specific to the arts (e.g., number of hours devoted to daily practice, artistic accomplishments, special schools attended, etc.). The latter set of items was comprised of items that signifi-

---

[1] This research was supported by Grant OEG-8-9-540215-4004 (010), Office of Education, U.S. Department of Health, Education, and Welfare.

[2] Requests for reprints should be sent to Lawrence R. James, who is now at the Institute of Behavioral Research, Texas Christian University, Fort Worth, Texas 76129.

cantly predicted creativity and other performance criteria in the scientific and engineering studies but that had been rewritten to correspond to artistic endeavors. Nonart students answered only the first 200 items, while art students answered all 249 items. Data for academic grade point average, sex, age, and experience in the arts were also obtained.

Criteria

Multiple criteria were obtained for the art subjects for the following four art areas: music, dance, theater, and the visual arts. For each art subject, one art faculty member completed separate single-scale ratings (15-point scale with even alternatives defined) on (*a*) sensitivity, (*b*) motivation, (*c*) expression of self, and (*d*) potential; a three-item creativity checklist (one summed score); as well as four specific art-area checklists (corresponding to the music, dance, theater, and visual-art areas, with each checklist providing a summed score across six items), on which art faculty members rated only subjects in their own art area (i.e., each subject was rated on one checklist only). In order to provide behavioral relevance to these criteria (i.e., wording the measures in meaningful artistic terminology that was amenable to measurement), the rating and checklist criteria were constructed with the assistance of art faculty and students. An exception was the creativity checklist, which retained the form previously employed in research on scientists and engineers (Taylor & Ellison, 1967). Additional criterion data for the art subjects included self-reported number of awards and major art-area grade point average.

ANALYSES AND RESULTS

Prediction of Art versus Nonart Criterion

An item analysis was conducted to predict the art versus nonart dichotomous criterion on the combined art and nonart samples ($N = 813$). Randomly selected halves of the art and nonart samples were combined, and a double cross-validation design was employed. Only the first 199 general biographical inventory items were analyzed. The average of the two cross-validities against the art versus nonart criterion was .67 (Fisher z).

Relations among Criteria

The correlations between the four ratings, the creativity checklist, the four art-area checklists, and major art-area grade point average were all significant and ranged from .33 to .83. Although the presence of some halo, or a lack of discriminant validity is indicated, the significant correlations between the four ratings (sensitivity, motivation, expression of self, and potential) and the separate art-area checklists provided evidence that the ratings were pertinent measures of artistic performance within each of the four art areas. Generally low to moderate correlations between the number-of-awards criterion and the remaining art-area criteria indicated the sampling of different criterion domains. The pattern of generally low correlations between the art criteria and academic grade point average ($rs = .11$–$.54$, mean $r = .26$) implied that different subsets of abilities were required for successful performance in these two fields. Finally, the correlations between the art criteria and the remaining control scores (e.g., age, sex, and experience in the arts) were generally low and not significant. An exception was the number-of-awards criterion which correlated .36 ($p < .01$) with experience in the arts. Experience in the arts was partialed out of the number-of-awards criterion.

Prediction of Separate Art Criteria Across Art Fields

An item analysis was conducted on the biographical inventory responses of the art sample to predict the different criteria of artistic performance. A double cross-validation design was employed on approximately equal subsamples of art subjects ($ns = 250$ and 251). All 249 biographical inventory items were analyzed; however, average cross-validities were obtained separately for the first 200 biographical inventory general items, the 49 biographical inventory items specific to the arts, and all biographical inventory items combined.

Significant average cross-validities ($p < .05$) were obtained for all art criteria. The average cross-validities provided by the 49 art items (highest average cross-validities per criterion ranged from .30 to .54) and all items combined (.31 to .57) were generally higher than the average cross-validities provided by only the 200 general items (.28 to .50). For example, the biographical inventory scoring key constructed to predict the sensitivity rating from the 200 general items had average cross-validities of .31, .41, and .50, respectively, against the sensitivity rating, potential rating, and a composite (equally weighted) of the potential rating and the number of awards (with experience partialed out). The biographical inventory key specifically constructed to predict the sensitivity rating on the 49 art items had average cross-validities of .37, .47, and .54, respectively, against these same criteria. However, only 2 of a possible 12 comparisons (for 12 criteria) between average cross-validities for the scoring keys provided by the general items versus the specific art items reached significance (two-tailed test for nonindependently distributed correlations, $p < .05$). Therefore, strong evidence was not provided that specific art items were more predictive of art performance than general items.

Predicting Art Criteria with the Art versus Nonart Scoring Key

Because the subsamples of art subjects were preserved throughout the analyses, it was possible to determine the average cross-validities of the art versus nonart scoring key (constructed to predict the art versus nonart criterion) against the art criteria for the art sample. With one exception, the average cross-validities of the art versus nonart scoring key against the art criteria were significant (.14 to .53). However, comparisons of the validities for the art versus nonart scoring key and the scoring keys developed on the 49 specific biographical inventory art items to predict the art criteria demonstrated that the latter scoring keys had significantly ($p < .05$) higher average cross-validities in 7 out of 12 possible comparisons. Further, the art versus nonart scoring key had significantly $p < .05$) lower validities in 3 out of 12 comparisons with the scoring keys developed on the same biographical inventory general items to predict the art criteria.

DISCUSSION

The present study indicated the feasibility of conducting studies of artistic performance using common criteria and predictors across different art areas, such as music, dance, theater, and the visual arts. This is not to imply that studies within separate art areas or comparisons between areas would not yield higher validities. Additional criterion development and construct validation of art criteria are needed.

The average cross-validity received for predicting the art versus nonart criterion in this study (.67) was approximately the same as the results found in the Anastasi and Schaefer studies (1969; Schaefer & Anastasia, 1968). However, the art versus nonart scoring key was often not as valid against actual performance in the arts as biographical correlates specifically determined by keying biographical inventory items to predict artistic criteria. A tentative conclusion was that different biographical inventory scoring keys should be employed to predict an art versus nonart criterion and performance criteria in the arts; however, predictive rather than concurrent validation studies are required for more conclusive evidence.

REFERENCES

ANASTASI, A. *Psychological testing.* New York: Macmillan, 1961.

ANASTASI, A., & SCHAEFER, C. E. Biographical correlates of artistic and literary creativity in adolescent girls. *Journal of Applied Psychology,* 1969, **53**, 267–273.

BURKHART, R. C. *Spontaneous and deliberate ways of learning.* Scranton, Pa.: International Textbook, 1962.

SCHAEFER, C. E., & ANASTASI, A. A biographical inventory for identifying creativity in adolescent boys. *Journal of Applied Psychology,* 1968, **52**, 42–48.

TAYLOR, C. W., & ELLISON, R. L. Predictors of scientific performance. *Science,* 1967, **155**, 1075–1079.

(Received August 31, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 87–89

USE OF WEIGHTING AND DATA ON ALTERNATIVE CAREERS TO PREDICT CAREER RETENTION AND COMMITMENT[1]

GERALD W. McLAUGHLIN AND RICHARD P. BUTLER[2]
Virginia Polytechnical Institute *U.S. Military Academy, West Point, New York*

This study investigated the desirability of collecting data on the importance of job characteristics and anticipated satisfaction with these characteristics in an alternative career to predict career commitment and retention. Subjects were 470 graduates of the U.S. Military Academy. Six data bases were developed and evaluated in terms of several validity measures. Results indicated that data on alternative careers can be meaningful, but data on the importance of job characteristics produced little practical improvement.

The purpose of this study was to evaluate two alternative strategies in data collection to predict career commitment and retention. These strategies concerned the value of using the importance of specific job characteristics as weights, and the value of using data concerning anticipated satisfaction with the job characteristics in an alternative career.

The desirability of item weighting has obtained both positive and negative support. Studies by Glennon, Owens, Smith and Albright (1960), Froehlich and Wolins (1960), and Youngberg, Hedberg, and Baxter (1962) led to the conclusion that item importance was relevant to job satisfaction. On the other hand, work by Larsen and Owens (1965) and Mikes and Hulin (1968) failed to substantiate these results. Ewen (1967) found mixed results which led him to conclude that weight might be of value, but should be empirically evaluated prior to utilization.

The question of obtaining data on satisfaction with alternative jobs has received less attention, with the primary impetus coming from research done by Porter (1962). The adaptation of this concept supports the belief that a person's job satisfaction is not a simple function of satisfaction with aspects in the current job, but is also related to the anticipated satisfaction with those aspects in an alternative job. It is thus a focus on relative satisfaction rather than absolute satisfaction.

METHOD

Subjects were 470 U.S. Military Academy graduates from the classes of 1963 through 1966 who responded to a mailed questionnaire concerning job satisfaction (response rate = 60%). The four classes were about evenly represented; and 331 of the respondents were still on active duty in the Army.

The questionnaire contained a list of 31 job characteristics that were derived from prior studies by Maslow (1954) and Herzberg (1959). Respondents were asked to rate the importance of a job characteristic in selecting a career, their satisfaction with it in the military, and their actual or perceived satisfaction with it as a civilian. The response scale for the importance of a job characteristic in selecting a career ranged from 0 to 4 (not relevant to crucial). Satisfaction was obtained on a 5-point scale of −2.0 to 2.0 (very dissatisfied to very satisfied).

Two criteria measures were obtained. The first was a measure of retention, that is, those subjects still on active duty scored 1, and those resigned from the military scored 0. The second criterion was used only for those still on active duty, and consisted of a 7-point scale that measured commitment to a military career. This scale has been shown to have reliability and predictive ability (Bridges, 1967), with high commitment having a score of 7 and low commitment a score of 1.

Three basic models were employed in the analysis of the data to develop initial data bases. The first model included information only on satisfaction with the 31 job characteristics in a military career. This data base is the one typically collected in studies of this type and was therefore named the classical model. The second data base contained two responses to each item, one for military satisfaction and one for civilian satisfaction, and is viewed as an alternative model. The final data base also contained information from both areas of satisfaction but used as the final measure the difference between military and civilian satisfaction on each item. It is therefore labeled the difference model.

---

[1] Any conclusions in this report are not to be construed as official U.S. Military Academy or Department of the Army positions unless so designated by other authorized documents.

[2] Requests for reprints and a more complete explanation of methodology and results should be sent to Richard P. Butler, Office of Institutional Research, U.S. Military Academy, West Point, New York 10996.

Multivariate procedures of factor and cluster analyses were used to reduce each data base to a set of scores. The weights for importance were used to form a second set of scores for each data base by multiplying each datum in a score by its importance. This methodology produced six different sets of scores for each respondent: classical, alternative, difference, weighted classical, weighted alternative, and weighted difference. Twelve forward stepwise regressions were run, using each of the six sets of scores with both of the criteria. Within each criteria, *t* tests for correlations from correlated data were computed to compare the validities obtained from the various sets of scores.

TABLE 1

COMPARISON OF SIX MODELS ON THE TWO CRITERIA

Model 1 vs. Model 2	Correlations			t
	$r^a_{1,c}$	$r^b_{2,c}$	$r^c_{1,2}$	
In vs. out				
Difference vs. weighted difference	.52	.51	.99	1.87
Alternative vs. weighted alternative	.54	.52	.97	2.12*
Classical vs. weighted classical	.47	.45	.93	1.46
Classical vs. alternative	.47	.54	.85	3.08*
Difference vs. alternative	.52	.54	.91	1.22
Career commitment				
Difference vs. weighted difference	.63	.65	.98	2.50*
Classical vs. weighted classical	.58	.60	.97	1.95
Alternative vs. weighted alternative	.63	.64	.93	.26
Classical vs. alternative	.58	.63	.90	2.18*
Difference vs. alternative	.63	.63	.94	.46

Note. For $N = 470$ (for the in versus out criterion) and 331 (for career commitment), a t value greater than ±1.96 is significant at the .05 level.

[a] $r_{1,c}$ is the correlation between the first model in a given row and the criterion.

[b] $r_{2,c}$ is the correlation between the second model in a given row and the criterion.

[c] $r_{1,2}$ is the correlation between the two models in a given row.

* $p < .05$.

RESULTS

The multivariate procedures produced 7 cluster scores for both the classical and difference data bases, and 15 for the alternative data base. The internal consistencies of the sets of cluster scores, for each of the three models, ranged from moderate to high. Cronbach alphas varied from .34 to .85, with a median of .69 for the classical model; from .34 to .86, with a median of .73 for the alternative model; and from .55 to .82, with a median of .73 for the difference model. The .34 alphas are much below the next lowest alphas, which were .61 for classical and alternative. Item weighting produced lower measures of internal consistency in 17 out of the 20 alphas which changed ($z = 3.1$, $p < .05$). This measure was significantly more than the 10.0 expected by chance.

The correlations between the models and the two criteria, and the tests of significance to compare the validities of the models are shown in Table 1. It should be noted that for the behavioral "in versus out" of service criterion, the alternative model was a significantly better predictor than the weighted alternative and classical models. For the career commitment criterion, the weighted difference and alternative models were significantly better predictors than the difference and classical models, respectively.

DISCUSSION

In terms of the efficiency of the final regression models, the validities for the weighted models were lower for the on active duty versus not on active duty criterion, and higher for the commitment criterion. The *t* values showed the differences to be significant for two of the six comparisons. However, the high correlations between the weighted and unweighted scores within a given data base enhanced the chance of finding small differences to be significant. It, therefore, appears that the differences of .02 in the multiple correlations that occurred have very little practical significance.

The use of the alternative data base produced significant improvement over the classical data base for the unweighted models for both criteria. However, the model for the alternative data did not make any significant improvement when compared to the difference model. Similar results were found for the parallel analyses with weighted data bases.

In conclusion, the results of this study fail to support the weighting of job satisfaction items by their importance when a behavioral, dichotomous criterion is used. There may be some slight benefit in applying weights when employing a criterion that produces continuous data. However, the benefits that can be derived appear to be so slight as to be more than offset by the added time and cost involved in collecting item weights. This study does, however, support the need to obtain job characteristic satisfaction on both the current and alternative job. While the decision to include this data as a difference measure for each, or to include each response separately might have theoretical interest, it did not affect the efficiency of the final regression equation.

REFERENCES

BRIDGES, C. F. *Research study on military career commitment.* West Point, N.Y.: U.S. Military Academy, Office of Institutional Research, 1967.

EWEN, R. B. Weighting components of job satisfaction. *Journal of Applied Psychology,* 1967, **51**, 68–73.

FROEHLICH, H. P., & WOLINS, L. Job satisfaction as need satisfaction. *Personnel Psychology,* 1960, **13**, 407–420.

GLENNON, J. R., OWENS, W. A., SMITH, W. J., & ALBRIGHT, L. E. New dimensions in measuring morale. *Harvard Business Review,* 1960, **38**(1), 106–107.

HERZBERG, F., MAUSNER, B., & SNYDERMAN, BARBARA. *The motivation to work.* New York: Wiley, 1959.

LARSEN, J. M., JR., & OWENS, W. A. The assignment of job attitude items to subscales. *Journal of Applied Psychology,* 1965, **49**, 172–181.

MASLOW, A. H. *Motivation and personality.* New York: Harper, 1954.

MIKES, P. S., & HULIN, C. L. Use of importance as a weighting component of job satisfaction. *Journal of Applied Psychology,* 1968, **52**, 394–398.

PORTER, L. W. Job attitudes in management: I. Perceived deficiencies in need fulfillment as a function of job level. *Journal of Applied Psychology,* 1962, **46**, 375–384.

YOUNGBERG, C. F. X., HEDBERG, R., & BAXTER, R. Management action based on one versus two dimensions of a job satisfaction questionnaire. *Personnel Psychology,* 1962, **15**, 145–150.

(Received September 25, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 90–92

PREDICTORS OF MANAGERIAL CAREER EXPECTATIONS[1]

MICHAEL L. MOORE[2]
Michigan State University

EDWIN MILLER
University of Michigan

JOHN FOSSUM
University of Wyoming

Relationships among managers' current managerial level, age, perceived position-learning time, level of formal education, and their career expectations were investigated. Results from a sample of 203 managers from five plants in one corporation indicated the current age of the manager was the most potent predictor of his career expectations.

At any given point in a manager's career, he is likely to have expectations regarding the level he will ultimately achieve in his company. This expectation is derived from such self-perceptions as his relative qualifications, opportunities to use his expertise, performance to date, and views of internal company politics. In addition, his expectations of promotability may be related to the length of time he perceives that it should take him to learn jobs enroute to his expected career endpoint (Glickman, Hahn, Fleishman, & Baxter, 1968, p. 25).

Isolating the social learning variable, Moore (in press) found support for Jennings's (1967) belief that most managerial jobs can be learned in two years or less. Moore found a mean perceived position-learning time of about 19 months with a mode of one year. Position-learning time is defined as the amount of time reported as necessary to learn the work responsibilities and duties defined by the organization's position description in terms of functions, hierarchical level, liaison, decision making and supervision.

The present study examines the relationships between perceived position-learning time, age and educational variables, and a measure of career expectations, to determine whether shorter perceived position-learning time or other factors were most related to career-expectation perceptions.

METHOD

Moore's sample provided data from 203 managers above line-foreman levels representing all functional areas in five plants in two different geographical regions of one manufacturer of consumer durables.

The respondents to this study were all males; 198 were married, 60.1% were less than 45 years of age and 83.7% were under 50. Most (82.2%) had had education varying from some college to graduate work, that is, demographically typical of managerial personnel in large organizations.

The perceived-learning time measure was an open-ended interval scale. Instructions were as follows: "How long has it taken or will it take you to learn your present job (from the time you began on the job)?"

The manager's estimate of the level of management expected to be achieved in his career was tapped by the item: "How high in the company do you expect to be promoted in your career: (a) president, (b) vice-presidential level, (*c*) senior management, at the divisional management level, (*d*) upper middle management, such as managers of large functional departments, (*e*) middle management, (*f*) lower middle management, such as supervisors of sections of functions, or (*g*) lower levels of management, including plant general foremen and supervisors of small functional units?"

Using information on present position level, and the seven levels of management given, managers were assigned to high- or low-career expectations groups. High-expectations group members (about 32%) were those at levels four or five who aspired to level three or higher; and those at levels six and seven who aspired to level four or higher. All others (about 68%) constituted the low expectation group.

The criteria for setting up the high-career

---

[1] This research was carried out at the University of Michigan and was first presented in the senior author's doctoral dissertation, *Managerial Learning and Socialization,* University Microfilms, 1969. The author reiterates his earlier thanks to the managers who cooperated in this study.

[2] Requests for reprints should be sent to Michael L. Moore, School of Labor and Industrial Relations, South Kedzie Hall, Michigan State University, East Lansing, Michigan 48823.

TABLE 1

Managerial Positions by Career Expectations Item Level

Title	Career expectations	
	Level	Proportion of salaried population
President	1	N.A.[a]
Vice president	2	N.A.[a]
Senior management	3	.001
Upper middle management	4	.008
Middle management	5	.024
Lower middle management	6	.04
Lower levels of management	7	.15

[a] N. A. = not applicable.

expectation group were derived from the firm's managerial manning tables. Table 1 shows the approximate percentage of salaried jobs at each final career expectation level. Table 1 masks the fact that in each level two through five, there are opportunities for three to four promotions, and thus those levels do not represent equal career intervals.

Initial analysis sought to determine whether a difference existed in perceived position-learning time between high and low expectation groups. Then, a multiple regression analysis of education (four levels), age (six levels), and position-learning time was run against the dichotomous career expectations variable. Both age and education were included as a set since all study data were categorized for these variables. One category of each was not included in the regression equation to avoid linear dependency and, hence, this category becomes part of the constant term (Wonnacott & Wonnacott, 1970, p. 75).

Results

Managers in the high expectations group ($n = 65$) had a significantly shorter mean perceived-position-learning time, 1.3 years, as compared to 1.72 years for the low expectations group ($n = 137$; $t = 2.83$, $p < .01$). Since the expectation group assignment depends on the manager's present position and his expected future promotions, variables such as past performance may not be reflected in expectations. Further, it is possible that variables such as present age and educational achievement level might be as highly related to career expectations as perceived position-learning time.

If the manager sees his educational background and current age as positive or negative factors in terms of his promotional possibilities, then these variables may predict career expectations as well as perceived position-learning time. Further, if learning time, in a multiple regression equation did not turn out to be a significant predictor, then one could say that its variance was accounted for by other variables. These possibilities were treated in the analysis illustrated in Table 2, which presents zero-order correlations between current managerial level, position-learning time, career expectations and various age categories and levels of educational achievement. Several trends and relationships can be observed.

Education. As might be expected, managers with high career expectations tend to be better educated. For those respondents who had only completed high school or some college, significant negative relationships were found. For those who had a bachelor's degree or had done some postgraduate work, the relationships were significant and positive.

Education and learning time. Surprisingly, the respondent's level of education accomplishment displayed an ambiguous relationship with respect to position-learning time. Only graduate education was inversely related to perceived learning time which could mean that effects on learning begin to be noted only when formal education

TABLE 2

Correlations between Background and Perceptual and Organizational Variables

Variable	*n*	Correlations		
		Current managerial level	Position-learning time	Career expectations
Current managerial level		—	.11	.13
Position-learning time		.11	—	−.14*
Educational achievement				
High school education	34	−.25**	−.02	−.23**
Some college education	61	.01	.08	−.26**
Bachelor's degree	83	.04	.02	.28**
Graduate education	25	.21**	−.15*	.25**
Age				
20–29	18	−.14*	−.15*	.25**
30–34	35	−.05	.10	.44**
35–39	27	−.02	−.08	−.07
40–44	42	−.13	−.05	−.08
45–49	48	.19**	.12	−.14*
50–54	25	.08	.23**	−.30**
Career expectations		.13	−.14*	

Note. $n = 203$.
* $p < .05$.
** $p < .01$.

reaches graduate level. Perhaps level of educational accomplishment is too gross a variable when related to position-learning time, and, in the future, more refined measures of education should be incorporated.

Age. Age proved to be a meaningful variable with respect to career aspirations. For managers ranging in age from 20 through 35, a relatively strong positive relationship existed for the two age categories and career expectations. For managers 35 years or older, the relationship was increasingly negative.

Thus, one can conclude that those individuals who are younger and college educated anticipate greater numbers of promotions and shorter periods of time spent in the steps up the managerial ladder.

To examine these data more closely, a multiple regression was run with career expectations group membership as the dependent variable and position-learning time, age, and educational levels as shown in Table 2 as independent variables. The multiple correlation was .62 ($F = 10.99$, $df = 11/191$, $p < .01$). However, only the two lowest age levels had significant regression coefficients. Age 20–29 was associated with high career expectations (t of beta $= 3.07$, $p < .01$) as well as age category 30–34 (t of beta $= 3.96$, $p < .01$). Thus it is likely that career expectation levels found in this study are largely a function of age, and not position-learning time.

DISCUSSION

A fundamental avenue of inquiry centers on the determination of why younger, better educated managers have higher career expectations than their older colleagues. Was there something in the work environment, the experience of older managers, and the company reward system which tended to work in a socialization perspective to dampen the career expectations of those managers who were over 34 years old? Perhaps age 35 represented that point in a manager's career when he "comes of age" and then more accurately assesses his career and the probabilities of accomplishing his goals. Or is it that after age 35, managers become cynical about life and their careers? Another possibility is that the organization's policy in personnel decisions is to prefer younger men.

In terms of path-goal motivation considerations (Evans, 1970; Georgopoulous, Mahoney, & Jones, 1957) it may be advantageous to a firm to develop means whereby performance can be shown to lead to faster or alternative paths to promotions. Learning a broad range of positions and responsibility areas may also be strengthened as a path to promotions. On the other hand, the older, less well-educated manager who does not expect rapid promotions may be much more concerned with other types of rewards. It might be worthwhile for a firm to trace out the perceived relationships between performance and promotions for the manager who is older and has not graduated from college.

Societal and labor market factors may also have been factors in the high expectations of younger managers in that they grew up after World War II and experienced an affluent society and a strong demand for technical and managerial talent throughout the 1960s. Without question, considerable research using socialization models (Brim & Wheeler, 1966; Brookover & Gottlieb, 1964), must be launched into an inquiry of why managers' career expectations change.

REFERENCES

BRIM, O., & WHEELER, S. *Socialization after childhood.* New York: Wiley, 1966.

BROOKOVER, W., & GOTTLIEB, D. *A sociology of education.* (2nd ed.) New York: American Books, 1964.

EVANS, M. G. The effects of supervisory behavior on the path goal relationship. *Organizational Behavior and Human Performance,* 1970, **5**, 277–298.

GEORGOPOULOS, B. S., MAHONEY, G. M., & JONES, N. W. A path-goal approach to productivity. *Journal of Applied Psychology,* 1957, **41**, 345–353.

GLICKMAN, A. S., HAHN, C. P., FLEISHMAN, E. A., & BAXTER, B. *Top management development and succession: An exploratory study.* New York: Macmillan, 1968.

JENNINGS, E. E. *The mobile manager.* Ann Arbor: University of Michigan, Bureau of Industrial Relations, 1967.

MOORE, M. L. Superior, self, and subordinate differences in perceptions of managerial learning times. *Personnel Psychology,* in press.

WONNACOTT, R. J., & WONNACOTT, T. H. *Econometrics.* New York: Wiley, 1970.

(Received October 24, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 93–95

WORK VALUES OF THE CHRONICALLY UNEMPLOYED [1]

DONALD J. SEARLS [2]
Colorado State University

G. N. BRAUCHT
University of Denver

R. W. MISKIMINS
Colorado State University

A scale of specific work values, the Work Requirements Rating Scale, was constructed and administered to supervisory, hourly wage, and chronically unemployed groups. The supervisory subjects were shown to differ from the other two groups in the high value they placed on basic work behaviors. The chronically unemployed and worker groups were found not to differ in any measureable way on work values. These and other findings of the study suggest that the scale used in this study has considerable promise as a means to assess work values.

Industry experiences frustration when it attempts to retain, in productive work, persons who have been chronically unemployed. These efforts might be enhanced by a better understanding of the work values of that group. A review of literature indicates that studies of work values are generally based on measurement of the importance attached to jobs in general (e.g., Super, 1957, 1962); many of these investigations focused on the relationship of work values and the Protestant ethic (e.g., Blood, 1969; Wollack, Goodale, & Wijting, 1971).

The primary purpose of this research was to develop a scale that would allow an examination of specific and commonly required on-the-job behaviors. The secondary objectives were to study the importance attached to those behaviors by supervisors, hourly wage employees, and economically disadvantaged unemployed individuals and, to examine sociocultural and personality characteristics of those three groups.

METHOD

Development of a scale of work values began with interviews of supervisors and personnel managers of seven large companies representing manufacturing, retail, and service businesses. The interviews were directed so as to identify those specific work behaviors which were conducive to either promotion or termination of hourly wage employees and resulted in an initial pool of 65 such behaviors. This list was culled to minimize ambiguity and item overlap. The 47 items that were retained formed the content of the Work Requirements Rating Scale and were found to fit into five logical subscales: (*a*) Personal Habits (11 items)—for example, "The worker has a sloppy or unkempt appearance"; (*b*) Supervisor Relations (7 items)—for example, "The worker sometimes argues with his supervisor"; (*c*) Employee Relations (7 items)—for example, "The worker has been unable to make friends with co-workers"; (*d*) Company Relations (10 items)—for example, "The worker has occasionally taken home inexpensive company material"; and (*e*) Work Habits (12 items)—for example, "The worker has a higher rate of absenteeism than most employees."

Each item in the Work Requirements Rating Scale was in a five-point Likert-type format. The scale listed five subscale scores and a total sum of work values. Subjects were asked to rate the importance of each item in considering a line worker for promotion. In addition to this scale, all subjects were administered the Social Access Questionnaire, which measured components of the sociocultural environment and personality (in part from Jessor, Graves, Hanson, & Jessor, 1968), previously shown to be predictive of vocational adjustment (Berry, Miskimins, & Keil, 1970). Five aspects of the sociocultural environment assessed by this instrument were: (*a*) opportunity for deviance—the extent to which an individual has deviant alternatives to work available; (*b*) exposure to deviant role models—the amount of exposure a person has had to non-conforming persons and forces; (*c*) absence of sanction networks—the degree to which an individual lacks access to social agents capable of exerting pressures against deviant behavior; (*d*) religion—the degree to which religion plays an important part in an individual's life; and (*e*) deviant behavior—the extent to which a person admits to engaging in deviant behavior.

---

[1] This investigation was supported by the U.S. Department of Labor (Contract No. 82-08-72-07) and was conducted through the Colorado State University Experimental Laboratory. Organizations undertaking projects under government sponsorship are encouraged to exercise their own judgment freely. Therefore, points of view or opinions stated in this article do not necessarily represent the official position or policy of the Department of Labor.

[2] Requests for reprints should be sent to Donald J. Searls, Colorado State University Experimental Laboratory, 50 West 5th Avenue, Denver, Colorado 80204.

Five measures of personality assessed by the Social Access Questionnaire were: (*a*) delay of gratification—a person's tendency toward immediate gratification; (*b*) alienation—the degree to which a person feels alienated from society's mainstream; (*c*) internal-external control—the degree to which an individual feels that he is in control of his life; (*d*) time perspective—the extent to which one plans for the future; and (*e*) attitudes toward deviance—an individual's attitudes toward many kinds of socially defined illegitimate behavior.

Both the Work Requirements Rating Scale and the Social Access Questionnaire were administered to the following three subject groups: (*a*) first line supervisors ($n = 23$) employed at a large regional warehouse for over 10 years (presently with supervisory responsibility for 6 or more employees); (*b*) hourly wage workers ($n = 52$) employed at the same work setting for over two years; (*c*) chronically unemployed individuals ($n = 50$) selected because of long histories of vocational failure and no recent attempts to find employment.

RESULTS

Each of the measures, previously described, were subjected to a one-way analysis of variance, with the supervisor (S), hourly wage worker (W), and unemployed (U) groups serving as the levels of the independent variable. When *F*s reached statistical significance, *t* tests were computed for all possible pairs (S–W, U–S, W–U).

Supervisor-worker comparison. The supervisor ascribed considerably more importance to basic work behaviors than did the line worker, and in fact, the average supervisor placed a numerically higher value on every one of the 47 items. The results for the five subscales and total were:

> Personal Habits ($F = 4.61$, $p < .05$; and S–W $t = 4.06$, $p < .01$); Supervisor Relations ($F = 6.22$, $p < .01$; and S–W $t = 4.57$, $p < .01$); Employee Relations ($F = 2.86$, f $= ns$); Company Relations ($F = 4.76$, $p < .05$ and S–W $t = 4.03$, $p < .01$); Work Habits ($F = 6.66$, $p < .01$ and S–W $t = 4.72$, $p < .01$); Total ($F = 5.31$, $p < .01$ and S–W $t = 4.20$, $p < .01$ —all *F*s 2/120 *df*, and all *t*s 73 *df*).

The supervisory and hourly wage subjects did not differ on any measure of sociocultural background.

Chronically unemployed-supervisor comparison. The Work Requirements Rating Scale shows that supervisors placed a higher importance on work requirements than chronically unemployed. Statistically significant differences were apparent on four subscales and the total score (see *F*'s in previous section;

> Personal Habits U–S $t = 3.24$, $p < .01$; Supervisor Relations U–S $t = 4.00$, $p < .01$; Company Relations (U–S $t = 3.51$, $p < .01$); Work Habits (U–S $t = 4.36$, $p < .01$); Total (U–S $t = 3.68$, $p < .01$—*df* for all *t*s = 69).

In regards to sociocultural background, the supervisors were subject to more sanction networks that prohibited deviant behaviors ($F = 8.92$, 2/122 *df*, $p < .001$ and U–S $t = 2.07$, $df = 71$, $p < .05$) and also have less exposure to deviant role models ($F = 3.65$, 2/122 *df*, $p < .05$ and U–S $t = 2.66$, 71 *df*, $p < .05$).

Chronically unemployed-worker comparison. An important finding of this study showed that the chronically unemployed and successfully employed hourly wage subjects did not differ on any factor of the Work Requirements Rating Scale. They differed on one factor of the sociocultural environment—sanction networks against deviant behaviors were more infrequent among the unemployed (see previous *F*; W–U $t = 4.07$, $df = 100$, $p < .01$).

In regards to personality variables, the most striking differences were found between the unemployed and the other two groups. The unemployed felt more externally controlled ($t = 2.28$, $df = 100$, $p < .05$) and were less able to delay gratification ($t = 4.02$, $df = 100$, $p < .01$).

DISCUSSION

The results of this study suggest that the scale constructed for the measurement of specific on-the-job work behaviors has considerable promise as a means to study work values. Although these findings support the study's validity and indicate that future work with the Work Requirements Rating Scale is justified, further research is necessary to clarify the influence of factors, such as job tenure and increased responsibility on the development, maintenance, and changing of work values.

REFERENCES

BERRY, K. L., MISKIMINS, R. W., & KEIL, E. C. *Research manual of the manpower laboratory.* (Final Report: Applied programs in manpower development) Washington, D.C.: U.S. Department of Labor, April, 1970.

Blood, M. R. Work values and job satisfaction. *Journal of Applied Psychology*, 1969, **55**, 456–459.

Jessor, R., Graves, T. D., Hanson, R. C., & Jessor, S. L. *Society, personality and deviant behavior*. New York: Holt, Rinehart, & Winston, 1968.

Super, D. E. *The psychology of careers*. New York: Harper, 1957.

Super, D. E. The structure of work values in relation to status, achievement, interests, and adjustment. *Journal of Applied Psychology*, 1962, **46**, 231–239.

Wollack, S., Goodale, J. G., & Wijting, J. P. Development of a survey of work values. *Journal of Applied Psychology*, 1971, **55**, 331–338.

(Received September 5, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 96–98

EFFECT OF SEX, RELIGION, AND ETHNICITY ON OCCUPATIONAL STATUS PERCEPTION

A. GEORGE GITTER, JULIE ALTAVELA, AND DAVID I. MOSTOFSKY[1]

Boston University

A $2 \times 3 \times 3$ design (subject's sex and religion and job incumbent's ethnic group) with 30 repeated measures (occupations) was used to investigate the perceived magnitude of (*a*) ascription (race, social class, and other accidents of birth), (*b*) actual achievement (real performance), and (*c*) gilded achievement (appearance of performance simulated by prevarication, "throwing the bull," etc.) in 30 occupations. Results indicated that actual achievement was perceived to be far more important than either ascription or gilded achievement, and the incumbent's ethnic group significantly influenced this perception. No significant effects were reported for the subject's sex or religion.

Often an employee's or applicant's perception of an occupation can affect the organizational milieu and cause additional problems of career identity, rationalized underproductivity, and interpersonal tensions. Analyses of the status components inherent in numerous occupations suggest much experimental research, little of which has in fact been pursued.

A number of standard sociology texts (e.g., Broom & Selznick, 1963) discuss two different types of status; ascribed and achieved. Ascribed status is defined as a person's position in a particular system that results from characteristics or attributes over which he has no control, that is, family background, race, sex, etc. Achieved status, on the other hand, refers to the position a person holds within a system that results from characteristics or attributes over which he has control, that is, educational attainment, performance of specified role tasks, etc. Thus, in the case of ascribed status one's position within a particular system is independent of the incumbent's behavior, whereas in the case of achieved status the incumbent's position is dependent on behavior.

Recently, an alternative conceptual framework has been proposed that, rather than categorizing status as either ascribed or achieved, views any given status as including three components of achievement: (*a*) ascribed, (*b*) actual, and (*c*) gilded (Gitter & Mostofsky, 1972). In that framework *actual achievement* was defined as actual performance of occupational role tasks, and *gilded achievement* as appearance of performance simulated by prevarication, such as "throwing the bull," etc.

The present study investigated the relationship between sex of perceiver, religion of perceiver, and the ethnic character of the occupational status incumbent on the perceived importance of these three components of occupational status. Because of the extreme competition in the job market, coupled with a focused interest on the employability of blacks and women, the study incorporated within a single design all these variables of interest.

METHOD

Subjects were 350 white undergraduates at Boston University. Table 1 shows the breakdown by sex and religion. Subjects were randomly assigned to one of three groups following the design shown in Table 1. A list of 30 occupations was compiled for the study. This list was almost identical to the one used by Gitter and Mostofsky (1972) in a previous investigation of occupational status. It included selected occupational statuses from Hatt and North's (1947) classic "prestige of occupations" study.

Included in the list were 11 pairs of occupations, with each pair representing a "supervisory–nonsupervisory" dyad. An example of such a dyad is foreman of bricklayers–bricklayer; foreman of bricklayers represents a supervisory position, and bricklayer represents its nonsupervisory counterpart. These dyads were included to investigate differences in perceived magnitudes of status components between supervisory and nonsupervisory levels within the same occupation, as well as differences among different occupations.

A self-administered questionnaire containing a list of the 30 occupations was completed by each subject during regularly scheduled class periods. For each occupation, the subject was asked to estimate the relative importance of the three

---

[1] Requests for reprints should be sent to David I. Mostofsky, Boston University, Department of Psychology, 64 Cummington Street, Boston, Massachusetts 02215.

TABLE 1

EXPERIMENTAL DESIGN

Item	Subject																	
	Black						Jewish						WASP					
	Male ($n = 38$)			Female ($n = 76$)			Male ($n = 34$)			Female ($n = 80$)			Male ($n = 40$)			Female ($n = 82$)		
	RC	P	J	RC	P	J	RC	P	J	RC	P	J	RC	P	J	RC	P	J
n	11	15	12	20	24	32	15	9	10	19	24	37	9	16	15	17	29	36

Note. RC = Roman Catholic; P = Protestant; J = Jewish, WASP = White Anglo-Saxon Protestant.

status components (ascribed, actual, and gilded achievement); however, these three terms were replaced in the questionnaire by the terms *family background*, *work*, and *bull*. Early pilot data showed that when particular instances which adequately defined ascription, achievement, and gilding are used, the terms are more easily understood by more subjects and require less instruction or explanation when the more general construct label is presented by itself. Subjects were further instructed to assume that the incumbents of all 30 occupations, which they were about to rate, were of particular ethnic background. One hundred and fourteen subjects were asked to assume that the members of the 30 occupations, which they were to evaluate, were black (black target group); another 114 subjects were to assume that the occupational members were Jewish (Jewish target group), and an additional 122 subjects were to think of "people in general," (white Anglo-Saxon-Protestant target group). After the subjects completed the questionnaire, they were asked to rate each of the same 30 occupations (disregarding any consideration of ethnic group) on a status ladder of occupations in the United States.

RESULTS [2]

An initial analysis of the two nonmanipulated variables (sex and religion of the subject) showed them to be nonsignificant. The remaining analyses were concerned with the subject's estimate of the magnitude of a status component as affected by (a) ethnicity of occupational status component (black, Jewish, white Anglo-Saxon Protestant), (b) type of status component (bull, family, work), and (c) type of occupation.

Differences between Target Groups

The influence of ethnicity on status incumbent was examined in three separate sets of analyses in which bull, family, and work scores were independently analyzed.

A simple analysis of variance among the target groups was performed for each of the 30 occupations. The results indicated that ethnic group significantly influenced the perceived magnitude of the bull component for 19 of the 30 occupations. The 11 occupations, in which ethnic group was nonsignificant, tended generally to be nonprofessional positions at or near the bottom of the status ladder.

In addition to the 30 analyses of variance investigating the effect of ethnic group, a pooled analysis was performed, retaining ethnic group as the independent variable and each subject's bull scores that averaged over all 30 occupations as the dependent variable. Results of the pooled analysis clearly supported other findings ($F = 1.58$, $df = 60/634$, $p \leq .01$). Similar analyses were conducted for family and work scores in which ethnic group was significant in the analyses for 8 and 9 occupations, respectively.

The effect of type of status component was tested by an analysis of variance with repeated measures; each subject's ratings of each of the status components were used as three repeated measures for each of the 30 occupations. The results indicated significant differences ($F = 168.20$, $df = 29/10092$, $p \leq .001$); actual achievement had approximately three times as much to do with being a member of an occupation as either gilding or ascribed characteristics (as seen in the overall means). The size of the gild (bull) component relative to that of the ascribed (family) component, varied with the type of occupation, such that the ascribed component was relatively smaller and the gild larger when the occupations were either management or sales oriented, and vice-versa when the occupations were professional. When the results of the analysis were repeated using each of the three target groups separately, the findings were virtually identical to those obtained when using the sample as a whole.

---

[2] Detailed tables of the individual ratings are available from the authors on request.

In addition, statistically significant differences between supervisor and nonsupervisor were obtained in perceived magnitude of the bull ($F = 102.26$, $df = 1/348$, $p \leq .001$), family ($F = 18.04$, $df = 1/348$, $p \leq .001$), and work ($F = 15.01$, $df = 1/348$, $p \leq .001$) components. Highest bull scores were associated with white Anglo-Saxon Protestant incumbents and lowest with blacks, with both supervisor ($F = 9.09$, $df = 2/346$, $p \leq .001$) and nonsupervisor ($F = 4.45$, $df = 2/346$, $p \leq .05$). Ethnicity of target groups did not significantly influence the magnitude of either the family or work components.

DISCUSSION

Several implications can be derived from the results of this study. First, this study provides an empirical validation of a theoretical schema that views status as consisting of a number of components. It also supports the general findings in the classic study of occupational prestige by Hatt and North (1947). Second, from the methodological standpoint, the study represents an advance over the technique used in the Gitter and Mostofsky (1972) study, in which subjects rated the magnitude of all three components for each of the occupations.

Substantively, the findings add to the total social psychology of occupational status. Actual achievement (work) emerges as a factor approximately three times as important in the incumbency of an occupation as either ascription or gilding. In addition, the ethnic group of the incumbent was shown to influence differentially the amount of actual, ascribed, and gilding that is attributed to him.

Caution should be exercised in making unqualified generalizations from the results of this study, since the findings, at least in part, may be a function of the nature of the sample employed, as in this case, a sample composed of rather liberal minority-group conscious students from a large urban university. Future research in this area might take this consideration into account by sampling from other populations including nonstudent adults who are themselves incumbents of the occupation studied.

REFERENCES

BROOM, L., & SELZNICK, P. *Sociology*. New York: Harper & Row, 1963.

GITTER, A. G., & MOSTOFSKY, D. E. Perception of "gilding" in occupational statuses. *Journal of Vocational Behavior*, 1972, 2, 333–341.

HATT, P. K., & NORTH, C. C. Jobs and occupations: A popular evaluation. *Opinion News*, 9, September 1947, 3–13.

LINTON, R. Concepts of role and status. In T. M. Newcomb & E. L. Hartley (Eds.), *Readings in social psychology*. New York: Holt, 1947.

(Received September 14, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 99–100

FACTOR STRUCTURE FOR BLACKS AND WHITES OF THE JOB DESCRIPITIVE INDEX AND ITS DISCRIMINATION OF JOB SATISFACTION

PATRICIA CAIN SMITH,[1] OLIN W. SMITH, AND JAMES ROLLO

Bowling Green State University

The Job Descriptive Index was administered to civil service whites (217), blacks (107), and predominantly white bank employees (110). The purposes were to extend convergent and discriminant validity, to compare factor structures for blacks and whites, to search for a general factor, and to check the discriminability of items for three disparate groups. The 72 item scores were factor analyzed. Factors for five scales, work pay, promotion, supervision, and co-workers were found for two or three groups with the exception of the supervision scale which split into two factors for all groups. Item means discriminated among the three groups as predicted from their known situations.

The Job Descriptive Index measures job satisfaction on five scales (work, pay, promotion, supervision, and co-workers) and is in frequent use (e.g., Blood, 1969; Evans, 1969; Hulin, 1968, 1969; Hulin & Waters, 1971; Soliman, 1970; Waters & Waters, 1969).

The purposes of this study were: (*a*) to determine whether the convergent and discriminant validity that was reported by Smith, Kendall, and Hulin (1969) for industrial workers would hold for civil service workers; (*b*) to compare the factor structure for blacks and whites; (*c*) to test for a general factor as predicted by Herzberg, Mausner, Peterson, and Capwell (1957); (*d*) to determine item discriminability of the five scales for three different white-collar samples of workers.

In the study by Smith et al. (1969) the subjects were predominantly industrial workers, and nearly all were white. The subjects of this study were all of the employees of a civil service accounting division except for absentees (212 whites, 107 blacks), and 110 randomly selected bank employees (nearly all white) with duties similar to those of the accounting division. For the latter sample the investigators predicted higher satisfaction because of superior working conditions and fringe benefits. Management employees of the accounting division predicted lower satisfaction for civil service blacks than whites because of fewer promotional alternatives.

A principal components analysis of the 72 items (scree test of Cattell, 1966) with varimax rotations yielded seven factors for each group of subjects. The work scale for civil service whites was split into descriptive and evaluative factors. (These factors require further investigation.) Loadings for pay, promotion and co-workers showed remarkable consistency.[2]

The supervision scale for all three samples was split into two factors which have been tentatively identified as quality of supervision (acts) and personal characteristics of the supervision.

Although there was no general factor for neither blacks nor whites, two unidentifiable factors were found. The method of construction, the specific contents, and the balanced positive and negative wording of the Job Descriptive Index may account for the absence of a general factor. The scales of the Job Descriptive Index did not exhaust those factors which may prove to be useful in quantitative studies of job satisfaction. The bias of this study, however, was that the understanding of job satisfaction can be achieved more economically by the use of multiple variables rather than one general factor.

Factor loadings of the three groups were correlated (a valid but crude measure of similarity according to R. B. Cattell, personal communication). The median correlation along the diagonals (likenamed factors) was .75 with a range of .49–.88 (see Footnote 2). Off diagonal correlations were low. The correlations of the loadings

---

[1] The authors appreciate the advice of C. J. Cranny and D. Steinmann.

This research was supported in part by a grant from the General Electric Foundation to the senior author.

Reprint requests should be addressed to Patricia Cain Smith, Department of Psychology, Bowling Green State University, Bowling Green, Ohio 43403.

[2] Complete tables may be obtained from the senior author.

of the two supervision factors were low or near zero. With the exceptions of the split in the work scale for civil service whites and the splits in the supervision scale, the scales remained stable for these samples of subjects and extended the results of Smith et al. (1969) for convergent and discriminant validity.

Comparison of means of the items for the three groups showed all of the differences in total items to be significant ($p < .001$) by sign test. Bank employees were satified with more items than civil service whites as predicted by the investigators. Civil service whites exceeded civil service blacks as predicted by management.[3]

CONCLUSIONS

Identities of factors for blacks were the same as for one or more groups of whites. Identities of the pay, promotion, and co-worker factors were the same for all three groups as for the subjects of Smith et al. (1969). For civil service whites, two work factors resulted, the stability of which requires further study. The original supervision factor was divided into two factors for all three new groups. The estimates of factor-matching of six factors across the three groups ranged from medium to substantial with the median substantial. Neither a race factor for blacks nor a general factor was obtained.

In terms of means for scale items, the Job Descriptive Index discriminated among three groups of subjects. Bank employees had the highest job satisfaction followed in order by civil service whites and blacks. The generality of applicability of the Job Descriptive Index to both white and black clerical subjects seems to be established.

REFERENCES

Blood, M. R. Work values and job satisfaction. *Journal of Applied Psychology*, 1969, 53, 456–459.

Cattell, R. B. The scree test for the number of factors. *Multivariate Behavioral Research*, 1966, 1, 245–276.

Evans, M. G. Convergent and discriminant validities between the Cornell Job Descriptive Index and a measure of goal attainment. *Journal of Applied Psychology*, 1969, **53**, 102–106.

Herzberg, F., Mausner, B., Peterson, R. O., & Capwell, D. F. *Job attitudes: Review of research and opinion*. Pittsburgh: Psychological Services of Pittsburgh, 1957.

Hulin, C. L. Effects of changes in job satisfaction levels on employee turnover. *Journal of Applied Psychology*, 1968, **52**, 122–126.

Hulin, C. L. Sources of variation in job and life satisfaction: The role of community and job-related variables. *Journal of Applied Psychology*, 1969, **52**, 279–291.

Hulin, C. L., & Waters, L. K. Regression analysis of three variations of the two-factor theory of job satisfaction. *Journal of Applied Psychology*, 1971, **55**, 211–216.

Smith, P. C., Kendall, L. M., & Hulin, C. L. *The measurement of satisfaction in work and retirement*. Chicago: Rand McNally, 1969.

Soliman, H. M. Motivation-hygiene theory of job attitudes: An empirical investigation and an attempt to reconcile both the one- and two-factor theories of job attitudes. *Journal of Applied Psychology*, 1970, **54**, 452–461.

Waters, L. K., & Waters, C. W. Correlates of job satisfaction and job dissatisfaction among female clerical workers. *Journal of Applied Psychology*, 1969, **52**, 388–391.

(Received October 24, 1972)

---

[3] An overall F test shows the differences in scale means to be significant ($p < .01$). The analysis by items is presented as more meaningful to users of the Job Descriptive Index.

Journal of Applied Psychology
1974, Vol. 59, No. 1, 101–102

ROLE AMBIGUITY, LOCUS OF CONTROL, AND WORK SATISFACTION

DENNIS W. ORGAN[1] AND CHARLES N. GREENE

Graduate School of Business, Indiana University

The present study sought to extend the generality of previous studies (based on Rotter's internal-external locus of control measure) concerning interrelationships among locus of control, perceived ambiguity in the environment, and satisfaction. A sample of scientists and engineers responded to questionnaires containing Rotter's measure of locus of control, a measure of role ambiguity, and two measures of job satisfaction. Results indicated that locus of control was related to both role ambiguity and satisfaction and that locus of control provided a greater independent contribution to satisfaction than did role ambiguity. Implications for interpreting scores on role ambiguity measures are noted.

A number of recent studies have found that measures of role ambiguity—the individual's feelings of uncertainty about his organizational role obligations and/or the means to fulfill them—are negatively and significantly related to satisfaction expressed by the individual (Kahn, Wolfe, Quinn, Snoek, & Rosenthal, 1964; Rizzo, House, & Lirtzman, 1970). These findings and their supporting theory imply that role ambiguity is aversive and, further, that it is a product of the organization and its environment (e.g., extent of change or turbulence affecting the organization, degree to which job behavior is specified by formal rules and policies, clarity and consistency of organizational communication).

There is evidence from other areas of inquiry, however, that the possession of job-related and other types of instrumental information is related to the belief system of the individual. From studies based on Rotter's internal–external locus of control measure, there is evidence that the type of person who actively seeks out information of an instrumental nature is one who tends to believe he can control his own outcomes, rewards, or fates (Rotter, 1966; Seeman, 1963). For example, internals (persons who score highly in the internal direction on Rotter's instrument, and thus believe their outcomes are shaped by their own actions rather than luck or powerful others) tend to be better informed about their occupations (Valecha, 1972). If internals report less role ambiguity and also report greater work satisfaction, then the relationship between role ambiguity and satisfaction could be merely symptomatic of a more basic relationship between locus of control and satisfaction.

This research sought, therefore, to answer the following questions: (*a*) Do internals report less role ambiguity than externals? (*b*) Do internals also report greater work satisfaction than externals? (*c*) If the answers to (*a*) and (*b*) are both affirmative, is there a significant relationship between locus of control and satisfaction when controlling for the effects of role ambiguity? (*d*) Which variable—locus of control or role ambiguity—provides the greater independent contribution to variance in satisfaction?

METHOD AND PROCEDURE

The sample consisted of 94 senior scientists and engineers employed in the research, development, and engineering divisions of a large manufacturer of electronics equipment, components, and supplies. Each participant filled out a questionnaire booklet which contained the following instruments: (*a*) an 11-item instrument developed by Rizzo, House, and Lirtzman (1970), used to tap subjects' perceptions of role ambiguity; (*b*) Rotter's (1966) Internal–External Locus of Control scale, which assesses the extent to which the respondent believes his outcomes in life are shaped by his own behavior rather than luck or powerful external agents; (*c*) an index of work satisfaction (satisfaction with the actual content of the job), consisting of five items drawn from Stogdill's (1965) job expectation questionnaire; and (*d*) a measure of general job satisfaction (satisfaction with the organization and its management), composed of 12 questions drawn from Stogdill's (1965) job description questionnaire.

Approximately half of the questionnaire booklets were completed at the work site and returned immediately. The remainder were returned by office mail within 10 days of administration.

---

[1] Requests for reprints should be sent to Dennis W. Organ, Graduate School of Business, Indiana University, Bloomington, Indiana 47401.

RESULTS AND DISCUSSION

Locus of control (scored in the direction of increasing perceived external locus of control) correlated significantly with role ambiguity ($r = .42$, $p < .01$), work satisfaction ($r = -.36$, $p < .01$), and general job satisfaction ($r = -.27$, $p < .01$). Role ambiguity also significantly related to work satisfaction ($r = -.30$, $p < .01$), but not to general job satisfaction ($r = -.14$, $p > .10$). It seems possible, then, that role ambiguity may elicit negative attitudes that are specific to the task and work behavior, but do not generalize to the individual's overall attitudes concerning the organization.

The partial correlation between locus of control and work satisfaction—controlling for role ambiguity—was also significant ($r = .27$, $p < .01$). On the other hand, the partial correlation between role ambiguity and work satisfaction—controlling for locus of control—was not significant ($r = .18$, $p > .10$). Thus, it appears that locus of control provides a greater independent contribution to differences in work satisfaction than does role ambiguity.

One further finding of interest emerged from a median split of the sample based on locus of control scores. For the subsample scoring in the direction of internal locus of control ($n = 45$), the correlation between role ambiguity and work satisfaction was significant ($r = -.29$, $p < .01$). However, among those scoring above the median on locus of control (i.e., in the external direction, $n = 49$), the relationship was nonsignificant ($r = -.19$, $p > .10$). This suggests that role ambiguity is aversive primarily to those who are frustrated in the active attempts to secure job-related information.

The results, when coupled with Valecha's (1972) findings that internals have more knowledge of the world of work (as measured by a 28-item occupational information test), support Rotter's (1966) hypothesis that "the individual who has a strong belief that he can control his destiny is likely to be more alert to those aspects of the environment which provide useful information for his future behavior." Thus, measures of role ambiguity bear an uncertain relationship to objective characteristics of the organizational environment, since they were also tapping the respondent's belief system and the instrumental value he placed on information.

Furthermore, the results would seem to support conclusions drawn by Korman (1971). Korman, whose study involved college and university students, found environmental ambiguity to be far less important as a determinant of satisfaction than the control aspects of the environment, that is, the extent to which students felt they were expected to exert self-control in student life. While the meaning of control as defined and measured by Korman differs somewhat from Rotter's conceptualization, both his study and the present research indicate that organizational ambiguity may not necessarily have the negative properties that some writers have attached to it.

Obviously, any conclusions drawn from the present study must be tempered by the correlational and self-report characteristics of the data. It is quite plausible, as Kahn et al. (1964) have implied, that role ambiguity is an antecedent of (or at least influences) perceived locus of control; or that both variables are concomitant effects of age, tenure, or other independent variables

REFERENCES

KAHN, ROBERT L., WOLFE, DONALD M., QUINN, ROBERT P., SNOEK, J. DIEDRICK, & ROSENTHAL, ROBERT A. *Organizational stress: Studies in role conflict and ambiguity*. New York: Wiley, 1964.

KORMAN, ABRAHAM K. Environmental ambiguity and locus of control as interactive influences on satisfaction. *Journal of Applied Psychology*, 1971, **55**, 339–342.

RIZZO, JOHN R., HOUSE, ROBERT J., & LIRTZMAN, SIDNEY I. Role conflict and ambiguity in complex organizations. *Administrative Science Quarterly*, 1970, **15**, 150–163.

ROTTER, J. B. Generalized expectancies for internal vs. external control of reinforcement. *Psychological Monographs*, 1966, **80**(1, Whole No. 609).

SEEMAN, M. Alienation and social learning in a reformatory. *American Journal of Sociology*, 1963, **69**, 270–284.

STOGDILL, RALPH M. *Manual for the job description and job expectation questionnaire*. Columbus: Bureau of Business Research, Ohio State University, 1965.

VALECHA, GOPAL K. Construct validation of internal–external locus of reinforcement related to work-related variables. *Proceedings, 80th Annual Convention of the American Psychological Association*, 1972, **7**, 455–456.

(Received October 12, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 103–104

PERVASIVENESS OF THE RELATIONSHIP BETWEEN INTRINSIC-EXTRINSIC NEEDS AND PERSISTENCE AT WORK

ELCHANAN I. MEIR[1] AND AZY BARAK

Tel-Aviv University

A total of 1,027 persons from 10 different occupations responded to inventories of intrinsic and extrinsic needs. While persistence at work was positively correlated with intrinsic needs, no correlation was found with extrinsic needs. These findings were generally consistent across occupations.

In a previous study (Meir, 1972) that used 288 dentists, 31 nurses, 20 policewomen, 30 social workers and 35 youth counselors as subjects, persistence at work was positively correlated with intrinsic needs while no correlation was found with extrinsic needs. It was consequently conjectured that these findings may also apply to women, and perhaps also to men, engaged in other occupations.

METHOD

The relationship between intrinsic and extrinsic needs on the one hand and persistence at work on the other was investigated in five more occupations—teachers, special education teachers, laboratory workers, psychiatric team workers and building workers. Replication studies were also made of two of the five occupations already investigated previously: nurses and youth counselors. The research population by group and sex is given in Table 1.

Intrinsic–extrinsic need inventories were constructed separately for each occupation. Each inventory consisted of a list of 7–19 intrinsic needs and 8–17 extrinsic needs.

The needs were classified by 4–24 psychologists and only those on which at least 80% agreement was reached were included. The scoring of the inventories has been described in detail in Meir (1972).

In 15 groups the criterion for persistence at work was the proportion of the total number of months elapsed since graduation. In two groups—special education teachers and policewomen—the criterion was dichotomy persistence versus nonpersistence since graduation. The hypothesis for the intrinsic needs was one-tailed and for the extrinsic needs two-tailed. The hypothesis for the intrinsic needs was one-tailed because it may be assumed that the more one is interested in the work itself, the longer is his persistence in it. The hypothesis for the extrinsic needs was two-tailed: On the one hand, it could be assumed that since these needs are satisfied through the work there will be a positive relationship between the extrinsic needs and persistence at work. Yet, on the other hand, since there exists the possibility that extrinsic needs might be satisfied even better in other occupations, the relationship can be assumed to be negative.

RESULTS

Table 1 shows the correlations of intrinsic needs and extrinsic needs with persistence at work.

DISCUSSION

The pattern of correlations between intrinsic needs and persistence at work indicated a positive relationship between the two variables. Seventeen out of the 18 correlations were positive ($p < .00008$), about half of them above .30. The only negative correlation was among women teachers in special education, while in the mixed group the correlation is .08. The disparity between salary and work requirements, on one hand, and difficulties for teachers in special education to accept a job in other kinds of educational work because of the shortage of personnel in special education, on the other, may possibly account for the deviation from the general rule.

Of the 18 correlations between extrinsic needs and persistence at work, 7 (39%) were positive. This finding by itself implied that including all occupations together the parameter of the relationship between the two variables is close to zero. The correlations for male workers in the separate and in the mixed groups indicated the same relationships as for the women.

It was preferred to use several smaller samples of different occupations rather than one big sample of one occupation since in the latter case the results would be applicable only to the

---

[1] Requests for reprints should be sent to Elchanan I. Meir, Department of Psychology, Tel-Aviv University, Ramat-Aviv, Israel.

TABLE 1

CORRELATIONS OF INTRINSIC AND EXTRINSIC NEEDS WITH PERSISTENCE AT WORK BY OCCUPATION AND SEX

Occupation	Subjects		Correlation		Correlation type
	No.	Sex	Intrinsic needs	Extrinsic needs	
Dentists[a]	288	F	.34**	−.13*	Pearson
Nurses[a]	31	F	.30*	−.14	Pearson
Nurses	68	F	.83**	.79**	Pearson
Policewomen[a]	20	F	.73*	.39	Biserial
Social workers[a]	30	F	.24	−.14	Pearson
Youth counselors[a]	12	F	.46	−.28	Spearman
Youth counselors[a]	11	F	.47	−.15	Spearman
Youth counselors[a]	12	F	.40	.19	Spearman
Youth counselors	141	M & F	.32**	−.08	Pearson
Special education teachers	38	F	−.53	−.80**	Biserial
Special education teachers	29	M & F	.08	.17	Spearman
Teachers	35	F	.24	−.22	Pearson
Teachers	95	F	.17*	.09	Pearson
Teachers	51	F	.47**	.06	Pearson
Teachers	28	M	.42*	−.10	Pearson
Laboratory workers	25	F	.29	−.18	Pearson
Psychiatric team	27	F	.42*	.03	Pearson
Building workers	86	M	.13	−.21	Pearson

Note. The hypothesis for intrinsic needs is one-tailed and for extrinsic needs is two-tailed. Abbreviations: M = male, F = female.

[a] Reported in Meir (1972).

* $p < .05$.

** $p < .01$.

population engaged in that particular occupation and not in other occupations. Although the occupations investigated were not chosen at random from the world of occupations, the relationship found seems to be a fairly general phenomenon. If social desirability affected the subjects' responses, it would lower correlations between intrinsic needs and persistence at work because of the lower reliability. A longitudinal study may show whether an intrinsic needs inventory has a predictive value for persistence at work.

REFERENCE

MEIR, E. I. The relationship between intrinsic needs and women's persistence at work. *Journal of Applied Psychology,* 1972, **56**, 293–296.

(Received October 18, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 105–106

RELATIVE EFFECTIVENESS OF FOUR COMBINATIONS OF ORAL AND WRITTEN PRESENTATIONS OF JOB RELATED INFORMATION TO DISADVANTAGED TRAINEES

M. M. PETTY[1]

Manpower and Industrial Relations Institute
University of Alabama

Relative effectiveness of four different combinations of oral and written presentations of information was investigated. It was hypothesized that the treatments which had an oral presentation as the first presentation of the combination would be more effective in disseminating information than the treatments which had a written presentation as the first presentation. The criterion of effectiveness was a multiple choice test designed to measure the subjects' knowledge of the disseminated information. All eight a priori comparisons made on the basis of the research hypothesis were in the expected direction, but only three were statistically significant ($p < .05$).

The purpose of the present study was to investigate the relative effectiveness of four combinations of different orders of presentations of oral and written information. The four combinations were oral and oral, oral and written, written and oral, and written and written. It was hypothesized that the treatments which had an oral presentation as the first presentation of the combination would be more effective in disseminating the information than the treatments which had a written presentation as the first presentation.

The development of the research hypothesis was influenced by previous research (Day & Beach, 1950) which had indicated that: (*a*) meaningful, familiar material was presented more efficiently orally; (*b*) the relative effectiveness of the visual presentation increased as the material became more difficult; and (*c*) the higher the ability level of the receiver, the greater the advantage of a visual presentation. Travers (1970), however, believed that the latter effect could be attributed to the fact that those who obtain the highest scores on such receiving tasks tend to be the best readers.

McCormick's (1964) finding that the use of two modalities was not systematically preferable to the use of one modality (where the amount of exposure was constant) focused attention on the order of presentations of oral and written information. The simplicity and saliency of the messages used in this study and the fact that the subjects were disadvantaged also influenced the nature of the research hypotheses (i.e., the predicted superiority of the oral presentation).

METHOD

Subjects

The subjects were 73 trainees who were attending a training program for the disadvantaged in Oak Ridge, Tennessee. The average age of the subjects was 22.3 years, 37% were school dropouts, and 49% were from minority groups. Six industrial skills (e.g., drafting, electronics) were taught as well as remedial basic education.

Treatments

Since it would not have been practical to have simultaneous presentations in the combinations where only one modality was used, an interval of 24 hours was provided between all presentations. Each of the four treatments was administered to four intact groups of subjects, that is, the experiment was divided into four phases with each group receiving a different treatment for each phase. The order of presentations of the treatments was counterbalanced to eliminate any systematic variance that might be due to a practice effect. The information was disseminated by the subjects' math instructors in order to maintain the normal atmosphere of the training program.

The first presentation of each treatment was given on a Monday while the second presentation was given the following day. A week after the first presentation of each treatment the subjects were tested on the material that was disseminated during the previous week. All subjects were tested at approximately the same time in their respective classrooms. Immediately following the tests, the message of the next phase was distributed through the assigned methods. Four weeks were required to disseminate the four messages and administer the appropriate tests.

---

[1] Requests for reprints should be sent to M. M. Petty, Manpower and Industrial Relations Institute, University of Alabama, P. O. Box 2737, Alabama 35486.

The oral portions of the treatments were standardized by having the instructor of each class read the information word for word. In the written presentations, each subject was provided with an individual copy of the material to read in class. Different information was used for each phase of the study, but all subjects received the same information during any one phase.

Messages

Four different messages, one for each phase, were developed to be disseminated to the subjects. Care was taken to ensure that the messages would provide the subjects with valuable information that would be beneficial to them when they began applying for a job.

The messages were edited and approved by the reading teacher in order to ensure their readability. Also, an analysis of the reading difficulty in each message was performed with the aid of the Flesch Reading Difficulty Index (Flesch, 1948). The results of this analysis indicated that three of the messages were rated fairly easy while the fourth one was rated easy.

Dependent Variables

The criteria were four multiple choice tests, one for each message. Each test consisted of 13 questions with five alternatives for each question. The test questions were phrased in a straightforward manner and in language similar to that used in the original material.

Covariate

Due to the nature of this field experiment, the use of intact groups was required to maintain the natural environment of the training program. It was realized that the classes may have differed on some variable that may have confounded the experiment. The most obvious dimension on which they might differ was general intelligence. Therefore, it was decided that a measure of the subjects' general intelligence could serve as a useful covariate. The general intelligence test of the General Aptitude Test Battery (GATB) was selected for this purpose.

RESULTS AND DISCUSSION

The data for each message was analyzed with a 2×2 factorial analysis of covariance (Barr & Goodnight, 1972, p. 144). The two independent variables were the first presentation and the second presentation both with two levels, oral or written. The results of these analyses indicated that the regression of the criterion on the covariate was significant ($p < .05$) for three of the messages—sources of jobs, tests and applications, and the interview (Fs $= 6.96$, 11.84, and 11.22, respectively; and dfs $= 1/58$, $1/55$, and $1/52$, respectively). The only significant treatment effect was the first presentation for the message on tests and applications ($F = 5.49$, $df = 1/55$, $p < .01$).

Eight a priori comparisons of adjusted treatment means were made. These means were adjusted for differences in general intelligence by partialing out the covariance between the criterion scores and the GATB scores. All eight comparisons were in the expected direction but only three were significant ($p < .05$), that is, the oral and oral treatment was more effective than the written and oral treatment in disseminating the message on the interview ($t = 1.76$, $p < .05$) while the oral and written treatment was more effective than the written and written treatment in disseminating the messages on tests and applications and social security (ts $= 2.11$ and 2.23, respectively, $p < .05$).

The three statistically significant comparisons provide some support for the following supposition: Initially, the best way to distribute relatively easy, job-related information to disadvantaged trainees is an oral presentation. This finding appears to support the conclusions of Day and Beach (1950) and Travers (1970).

The primary implication of this study is that supervisors as well as trainers must be cognizant of the possible limitations of written communications which are disseminated to disadvantaged personnel. The results of this study also imply that if a written communication is to be distributed to disadvantaged personnel it should be preceded by an oral version of the same message.

REFERENCES

BARR, A. J., & GOODNIGHT, J. H. *A user's guide to the statistical analysis system.* North Carolina State University, August, 1972.

DAY, W. F., & BEACH, B. R. *A survey of the research literature comparing the visual and auditory presentation of information.* (Contract No. W33-039-ac-21269, E.O. No. 694-37) University of Virginia, November 1950. Cited by R. M. W. Travers, *Man's information system.* Scranton, Pa.: Chandler Publishing Co., 1970.

FLESCH, R. A new readability yardstick. *Journal of Applied Psychology,* 1948, 32, 221–233.

MCCORMICK, E. J. *Human factors engineering.* New York: McGraw-Hill, 1964.

TRAVERS, R. M. W. *Man's information system.* Scranton, Pa.: Chandler Publishing, 1970.

(Received October 24, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 107–109

PERSONALITY STABILITY AND ADJUSTMENT TO AN EXTREME ENVIRONMENT

JAMES N. BUTCHER[1] AND MICHAEL RYAN[2]

University of Minnesota

Personality characteristics of a group of Antarctic explorers were examined and successive administrations of the Minnesota Multiphasic Personality Inventory (MMPI) and Personality Research Form (PRF) were used to assess personality changes resulting from the wintering-in experience. Personality assessment suggests that the volunteers are generally better adjusted, more achievement oriented, more self-sufficient and serious minded than a control sample of normal males. No changes in personality profiles were found in retesting at midwinter and at the end of winter.

There has been a great deal of scientific interest in the exploration of the arctic and antarctic regions to gain a more detailed understanding of the planet on which we live. A number of expeditions, by many nations, have been sent to Antarctica in the last 25 years for purposes of charting, collecting samples of minerals, animal and plant life, and to make weather observations. Somewhat incidental to the general purposes of these expeditions has been an interest in studying the effects of arctic exploration on human adjustment. Individuals who participate in these prolonged expeditions to Antarctica are exposed to extreme climatic conditions and the threat of physical danger as well as to the psychological stress inherent in the long period of isolation characterized by sameness, boredom, and irritations occurring from being housed in cramped quarters.

Solitary voyagers and early polar explorers Bernicot, 1953; Byrd, 1938; Gibson, 1953; Merrien, 1954; Slocum, 1918) have reported personality or experiential disorders such as delusional and hallucinatory experiences resulting from isolation, lack of sleep, starvation and fear of the elements. For obvious reasons it has not been possible to acquire reliable personality data pertinent to evaluating psychological adaption under extreme environmental conditions. Most of the existing knowledge of these personality factors comes from anecdotal report.

In recent years, since the 1957 geophysical year expedition, it has been possible to obtain objective data relevant to personality and social behavior on individuals who are isolated in various size groups for long periods of time. These wintering expeditions are somewhat different from the earlier explorations in that (at least the American sponsored expeditions) a greater degree of creature comfort has been possible since the U.S. Armed Forces' bases are relatively permanent, thus making possible the storage of supplies and luxuries such as movies, whiskey, etc. Despite the fact that greater materialistic comfort is reasonably assured, the situation still presents isolation, boredom, potential hazard and the knowledge that return to the safety of home is not possible for some time.

A number of studies have examined problems in adjustment to the antarctic environment. Several investigators have reported a high incidence of sleep disturbance, depression, headaches, irritability, and other minor emotional problems (Gunderson, 1963, 1968; Gunderson & Nelson, 1963; Mullin, 1960; Nelson, 1965; Seymour & Gunderson, 1971).

The danger, hardship, and cold are not serious sources of stress according to Mullin (1960). He found that more stress resulted from: (*a*) the individual's adjustment to the group, (*b*) the sameness of the environment and, (*c*) the absence of usual sources of emotional gratification. All of the available studies focusing on emotional adjustment of personnel wintering in Antarctica report instances of adjustment problems and stress. However, there have been *no* documented psychiatric breakdowns of psychotic proportions (Nelson, 1965).

The present study is an investigation of the personality characteristics of a group of antarctic volunteers who wintered-in during the 1970 winter. It was the investigators' hope to contribute an objective personality study of well-functioning individuals living in a stressful environment to the scientific understanding of personality. The objectives of this study were to provide an (*a*) objective description of the men who volunteered (and were selected) for this difficult duty and (*b*) an assessment of personality change, if any, which may have taken place during the long antarctic winter.

---

[1] The authors wish to thank Marjorie Sullivan and Richard Weber for assistance in the data analysis.

Requests for reprints should be sent to James N. Butcher, Department of Psychology, University of Minnesota, Elliott Hall, Minneapolis, Minnesota 55455.

[2] The data for this study were collected by Dr. Ryan who served as station physician with the United States Navy.

METHOD

The subjects were 15 volunteers from a group of 21 scientists and naval support personnel, who wintered in Antarctica at South Pole Station in 1970. All subjects had undergone extensive screening which consisted of physical examination and interviews with a psychologist and a psychiatrist prior to deployment. The South Pole Station, a small underground facility, housed around 20 personnel through the 8.5-month period of isolation. During this wintering-in period (the base was entirely self-sustaining), the only contact with the outside world was by short-wave radio.

The Minnesota Multiphasic Personality Inventory (MMPI) and Personality Research Form (PRF) were administered to the volunteers (*a*) just after the station was closed for the winter (March, 1970), (*b*) at midwinter and, (*c*) just prior to the reopening of the station at the end of the antarctic winter (September 1970). The station physician administered the tests (see Footnote 2). The subjects were given the test materials and asked to complete the inventories and return them as soon as possible. They were requested not to discuss or collaborate with anyone in responding to the inventories. All materials were returned within two days.

The control subjects ($n = 30$ college males) were randomly selected from a group of 200 normal college males who had been administered the MMPI and Personality Research Form on two occasions eight months apart.[3]

RESULTS AND DISCUSSION

The first question of interest to this investigation concerns measured personality characteristics of the antarctic volunteers: How do these individuals differ from people in general? The MMPI profile of antarctic volunteers was one which resembled group mean profiles of nonpathological groups or normal subjects (Lanyon, 1968). To compare the volunteers with a sample of normals, their initial MMPI and PRF profiles were compared with those of a group of male college subjects.[4] When compared with this group of normals, the antarctic volunteers appear better adjusted in the following ways: (*a*) They showed significantly less anxiety as reflected in scales Psychasthenia and Anxiety ($F = 29.8$, $p < .001$; $F = 15.2$, $p < .001$, respectively). (*b*) Less bodily concern shown by lower scores on scales Hypochondriasis and Hysteria ($F = 6.3$, $p < .05$; $F = 5.9$, $p < .05$, respectively). (*c*) They appeared to be more conventional and less alienated as shown by lower scores on scales Psychopathic Deviate and Schizophrenia ($F = 4.4$, $p < .05$; $F = 8.4$, $p < .01$) than the younger college subjects. The greater elevation on the Ego-Strength scale ($F = 15.2$, $p < .001$) shown by the antarctic volunteers suggest that they are perhaps more able to withstand stress than the college males. These findings were consistent with the expectation that self-selection and the preliminary psychiatric screening operate to eliminate individuals who have lesser adaptive capabilities. The generally more stable or better adjusted profile of the volunteers may be related to the fact that the volunteers were somewhat older than the college males.

The personality differences between the volunteers and the college males on the PRF were informative. In general, the volunteers tended to describe themselves in a more favorable way than did college males (Desirability scale score difference $F = 8.8$, $p < .01$). This finding was consistent with the MMPI results previously reported in this study. The volunteers present themselves as more achievement oriented than the college males (Achievement scale score difference, $F = 9.7$, $p < .01$). This was not a surprising finding since a number of the volunteers were scientific professional individuals who had achieved both academically and professionally. The result added to the construct validity of the PRF achievement scale.

Two other PRF scale differences pointed to the more serious-minded and more self-sufficient attitudes of the volunteers. The volunteers produced lower Play scores (Play scale score difference, $F = 12.9$, $p < .001$) than the college males. This suggested less of an easy-going attitude toward life; for example, less participation in games, sports, or doing things just for fun by the volunteers than by the college males. The volunteers were lower on the Succorance Scale ($F = 4.6$, $p < .05$) which showed that the college males more frequently sought sympathy, protection, love, advice and reassurance of other people than did the volunteers.

---

[3] These data were collected by James Butcher and Paul Mauger in cooperation with Bethel College, St. Paul, Minnesota.

[4] A more complete report of the findings may be obtained by writing the first author.

Assessment of Personality Change

Analyses were performed on MMPI and PRF profiles which were obtained in pre- and post-winter testing sessions. The stability of the MMPI and PRF for both the volunteers and the college males was shown by the lack of significant differences in mean scores over an 8-month period of time. Only one score was significant at the .05 level out of 76 comparisons, a finding which was consistent with chance expectancy.

A second analysis was performed to compare difference scores between first and second MMPI and PRF administrations for the volunteers and the college males. For this analysis the squared difference between the first and second administrations on the 16 MMPI scales and the 22 PRF scales were used. In the 38 comparisons of mean difference scores between the volunteers and the college subjects only one difference reached significance. This can best be interpreted as a chance finding. The results of this analysis showed that MMPI profiles of antarctic volunteers did not change (in terms of elevation) as a result of the wintering-in any more than a control group of college subjects who had not undergone the experience.

Both of the previous analyses focused on changes in scale elevation and did not take into account scale configuration which is often considered more crucial to profile interpretation than elevation. In order to evaluate possible shifts in scale configuration, a third analysis was performed to assess possible changes in the profiles. The MMPI profiles of volunteers and control subjects were coded following the Welsh coding system. The first four code digits (the four most elevated scales) for both the experimental and control subjects were examined and the initial and retest configurations compared. Each subject was given a score value of one for each scale which shifted position out of the first four code-digit group of scales between the initial testing and retest. The results of this analysis indicated that scale configuration for volunteers and for college normals remained essentially the same. The volunteers all had only one or two code shifts with an average of 1.4 and the college students had between zero and four code shifts with most subjects having only one scale shift. The average code shift for the college subjects was 1.4 or identical to the average for the volunteers.

Thus both elevation and configuration for the volunteers and for the normal college men remained fairly constant between initial testing and retesting some months later regardless of the intervening events. Replication of these findings on another antarctic expedition could be valuable.

Midwinter Data

The MMPI and PRF were administered to the volunteers at midwinter since previous research had indicated that adjustment problems were most apparent at this point in the winter. No significant differences for any of the scales on the MMPI or PRF were obtained between the initial test and the midwinter retest. Although some of the volunteers experienced some symptoms of maladjustment such as headaches, depression, irritation, etc., these problems were not sufficient to alter deviantly their MMPI profiles.

REFERENCES

BERNICOT, L. *The Voyage of Anahita—Single-handed Around the World*. London: Rupert Hart-Davis, 1953.

BYRD, R. E. *Alone*. New York: Putnam & Sons, 1938.

GIBSON, W. *The Boat*. Boston: Riverside Press, 1953.

GUNDERSON, E. K. Emotional symptoms in extremely isolated groups. *Archives of General Psychiatry*, 1963, **9**, 362–368.

GUNDERSON, E. K. Mental health problems in Antarctica. *Archives of Environmental Health*, 1968, **17**, 558–564.

GUNDERSON, E. K., & NELSON, P. D. Adaption of small groups to extreme environments. *Aerospace Medicine*, 1963, **34**(12), 1111–1115.

LANYON, R. I. *Handbook of MMPI Group Profiles*. Minneapolis: University of Minnesota Press, 1968.

MERRIEN, J. *Lonely Voyagers*. New York: Putnam & Sons, 1954.

MULLIN, C. S. Some psychological aspects of isolated antarctic living. *American Journal of Psychiatry*, 1960, **117**, 323–325.

NELSON, P. D. Psychological aspects of antarctic living. *Military Medicine*, 1965, **130**(5), 485–489.

SEYMOUR, G. E., & GUNDERSON, E. K. Attitudes as predictors of adjustment in extremely isolated groups. *Journal of Clinical Psychology*, 1971, **27** (3), 333–338.

SLOCUM, J. *Sailing Alone Around the World*. London: Rupert Hart-Davis, 1918.

(Received October 25, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 110-112

USING VARIANCE AS A DISCRIMINATOR IN LIE DETECTION

HOWARD WAINER[1]
University of Chicago

GUNNAR T. GRUVAEUS
Princeton University

MICHAEL O. BLAIR AND NICHOLAS ZILL II
University City Science Center

Using galvanic skin response (GSR) data from an earlier study by Orne and Thackray, it was shown that variance of response was a suitable discriminator between liars and truthtellers when an individual's deception is detected on a single trial. Subjects who were liars differed from subjects who were truthtellers on all trials, not just on the ones in which they tried to deceive the experimenter. The discrimination between liars and truthtellers was facilitated by Gruvaeus and Wainer's method of cluster analysis.

The use of galvanic skin response (GSR) as a physiological means of lie detection has been under investigation for an extended period of time (Gustafson & Orne, 1963, 1964, 1965a, 1965b). One typical procedure has been to form two kinds of groups, liars and truthtellers, in which the liar group has been required to learn a number of words. These learned words (designated "hot") are then mixed in a presentation sequence with neutral words and presented to the two groups together. All subjects were asked to report whether or not each word presented was one they were required to learn. The liar group was told to respond with "No" for all words, and hence was requested to lie when a hot word was presented. The truthteller group, of course, always responded with "No." The GSR of each individual was continuously recorded. A more detailed description of this experimental procedure was found in Orne and Thackray (1967).

Two objectives were to determine which words were hot and which subjects were liars. As has been pointed out previously (Orne & Thackray, 1967), this is essentially a problem in signal detection, in which it is essential to isolate the signal (physiological response to hot words) from the noise (physiological response to neutral words).

One quick and easy way to increase the signal to noise ratio was to average across hot words. Orne and Thackray (1967) carried out this method and found that liars may be discriminated from truthtellers, even though there was considerable overlap of the two distributions. However, when averages were taken of the subjects as well as the words the overlap of liar and truthteller groups decreased considerably. This finding indicated that there were differences between the two groups that were discernable in maximum deflection, but as Orne and Thackray point out, this is of little use for discrimination of individual subjects.

The purpose of this study was to indicate how another quality of the GSR curve other than maximum deflection might be of some use. Observation of many individual response curves indicated that subjects who lie "jiggle around" more than truthful ones and, thus, it seemed appropriate to characterize this jiggle by each individual's response variance around his own mean. When this was done for each word it was found that a comparison could be made between each subject's jiggle with each other subject by computing a distance between every pair of subjects. This distance was a function of their variances. More precisely the distance between subject i and subject j was defined as d_{ij}, where

$$d_{ij} = | \log [s^2_i/s^2_j] | .$$

The log of the variance ratio was taken because the distribution of distances was typically very skewed and the log transformation placed it in a more sensible scale. The use of the absolute value of the log was to conform to the symmetry requirement of a metric ($d_{ij} = d_{ji}$). A distance matrix between the subjects for each word was then computed and clustered using Gruvaeus and Wainer's (1972a, 1972b) technique. This kind of cluster analysis yields ordered hierarchical cluster solutions. Two sample results are shown in Figures 1 and 2 for a "hot" and "cold" word respectively for 20 subjects. The numbers along the top are subject

---

[1] The authors thank Martin T. Orne who very generously allowed them to use his data.

Requests for reprints should be sent to Howard Wainer, Department of Psychology, University of Chicago, Illinois 60637.

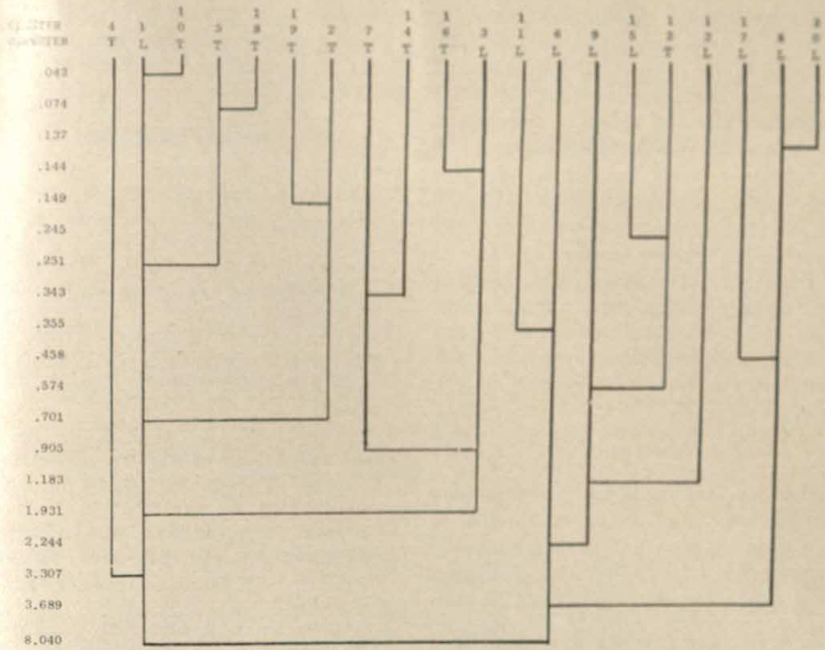


FIGURE 1. Ordered hierarchical cluster analysis of intersubject distances from a "hot" word. (L and T represent the subjects who are liars and truthtellers.)

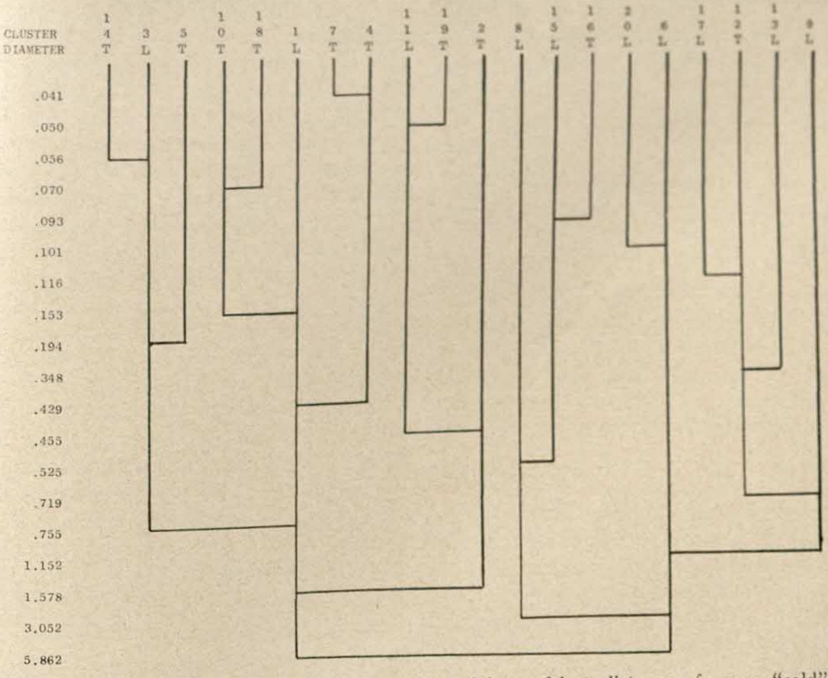


FIGURE 2. Ordered hierarchical cluster analysis of intersubject distances from a "cold" word. (L and T represent the subjects who are liars and truthtellers.)

numbers and the L and T represent whether the subject is lying or truthtelling, respectively. The numbers along the side are the cluster diameters. It is necessary to keep in mind that the *order* of the subjects across the figure is meaningful, that is, in Figure 1, for example, Subject 3 is more like Subject 11 than like Subject 20. This order implies that even though Subjects 3 and 11 are in different clusters Subject 3 is more like the subjects in the liar cluster than any of the other subjects in his cluster. Note also that the ability of the procedure to discriminate between lying and truthtelling subjects is quite good for both hot and cold words. The differences for the hot words seem to be somewhat larger. The two cluster solutions shown are quite typical for the 75 trials that have been analyzed in this way.

Although it was not necessarily the purpose of this study to propose a cluster analysis of intersubject distances as a way to determine who was lying, such a procedure may be conceivable. Rather, this study has been an attempt to indicate that by examining variance of response, large differences between lying and truthtelling subjects appear without having to resort to aggregating of the words or subjects. The fact that these differences showed up on cold words might be artifactual, caused by Orne and Thackray's (1967) differential treatment of groups in their study. Nevertheless, it is clear that accurate discrimination is possible in only one trial and that perhaps by concatenation of the trials the two groups can be distinguished.

REFERENCES

GRUVAEUS, G., & WAINER, H. MOCA: A 360/65 program for multiple ordered cluster analysis. *Behavioral Science*, 1972, 47, 490. (a)

GRUVAEUS, G., & WAINER, H. Two additions to hierarchical cluster analysis. *The British Journal of Mathematical and Statistical Psychology*, 1972, **25**, 200–206. (b)

GUSTAFSON, L. A., & ORNE, M. T. Effects of heightened motivation on the detection of deception. *Journal of Applied Psychology*, 1963, **47**, 408–411.

GUSTAFSON, L. A., & ORNE, M. T. The effects of task and method of stimulus presentation on the detection of deception. *Journal of Applied Psychology*, 1964, **48**, 383–387.

GUSTAFSON, L. A., & ORNE, M. T. Effects of perceived role and role success on the detection of deception. *Journal of Applied Psychology*, 1965, **49**, 412–417. (a)

GUSTAFSON, L. A., & ORNE, M. T. The effects of verbal responses on the laboratory detection of deception. *Psychophysiology*, 1965, 2, 10–13. (b)

ORNE, M. T., & THACKRAY, R. I. Group GSR techniques in the detection of deception. *Perceptual and Motor Skills*, 1967, **25**, 809–816.

(Received November 6, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 113-115

EFFICIENCY OF GSR DETECTION OF INFORMATION WITH REPEATED PRESENTATION OF SERIES OF STIMULI IN TWO MOTIVATIONAL STATES

ISRAEL LIEBLICH,[1] GIDEON NAFTALI, JOSEPH SHMUELI AND
SOL KUGELMASS

Hebrew University of Jerusalem, Jerusalem, Israel

The present study explored the possibility of reducing errors in the detection of information using the galvanic skin response (GSR), by employing a longer interrogation with repetitions of the same critical item. The experiment involved two groups of subjects, one tested under instructions intended to arouse high motivation to beat the machine; the other was tested under low motivation to beat the machine. It was found that it was possible to reduce considerably errors in information detection through a simple combination of GSRs obtained from 10 repetitions of the same interrogation sequence. Motivation did not produce any noticeable effect on the efficiency of detection of information.

One of the most important characteristics of the scientific investigation of the detection of information by psychophysiological means is the acceptance of some fallibility of the detection process, as an empirical fact. Thus attempts can be made to reduce error associated with the process, using methods that are generally known to be effective in reduction of error terms. It is clear by now that special state methods or magic new psychophysiological indices have *by themselves* little promise for dramatic improvement in accuracy, although they may help in providing additional information for the reduction of error terms. Three independent alternatives for the reduction of error terms in this area present themselves: (*a*) the simultaneous utilization and optimal combination of many psychophysiological indices (Kugelmass & Lieblich, 1968); (*b*) the use of larger alternative ensembles into which the critical items are inserted (Lieblich, Kugelmass, & Ben Shakhar, 1970); (*c*) the use of a longer interrogation which utilizes many critical items indicating the sought information (Lykken, 1959; Ben Shakhar, Lieblich, & Kugelmass, 1970). The present study explores a fourth alternative, which is a special case of (c), when only one critical item may be used to detect involvement with information: (*d*) the use of longer interrogation with repetition of the same critical item. Preliminary work (Nevo, 1968) has shown that the latter alternative might be feasible.

This study was intended to investigate the effects of repetition of series, in which the same critical item is embedded, on the probability of correctly identifying these critical items, through the use of changes in galvanic skin responses (GSRs). Since it was assumed that the repetition of the interrogation sequences would produce boredom, and since Gustafson and Orne (1963) showed that low motivation may reduce detection rates to chance level, it was decided also to manipulate motivation of the subjects along the lines of Gustafson and Orne (1963).

METHOD

Subjects

Fifty-eight students of the Hebrew University served as subjects in the reported study.

Instruments

The subjects were tested with a constant current (20 μA), GSR apparatus (Lieblich, 1966) coupled with a Brush mark II recorder. Beckman electrodes were used and were attached to the volar side of the second and fourth fingers of the left hand, using Beckman electrode paste.

Procedure

When each subject arrived at the laboratory, he was asked his first name. An interrogation sequence was constructed comprised of five names of which one was the first name of the subject. After the experimenter seated the subject and arranged the recording attachments, the experimenter entered the adjacent control room where he conducted the experiment. The interrogation sequences were read by an experimenter who did not know the subject's name, using an intercom. The subject was observed through a one-way screen. Two minutes were allowed for stabilization before starting. The subjects were randomly assigned to one of two groups, a high-motivation or a low-motivation group.

---

[1] Requests for reprints should be sent to Israel Lieblich, Department of Psychology, Hebrew University of Jerusalem, Jerusalem, Israel.

The instructions to the high- and low-motivation groups were very similar to those used by Gustafson and Orne (1963). Following the instructions the interrogation sequence began. The first name read was a buffer item followed by the sequence names. The questions were posed at ten second intervals. Ten repetitions of the names sequence were presented continuously. The order of the names was different for each repetition. In all, 51 names were presented to each subject.

RESULTS

In order to evaluate the effects of the high-motivation group instructions the subjects were interviewed immediately after the experiment. Each subject was asked if he had used any specific technique designed to beat the machine. Eight of the 28 high-motivation group subjects reported that they had not tried to use any specific means of interference with the polygraphic detection. For purposes of evaluating the effects of the motivation instructions, it seemed appropriate to analyze the results of the high-motivation group, both including and eliminating the eight subjects. The subjects in the high-motivation group who reported the use of a specific technique designed to interfere with the detection procedure were called high motivational countermanipulators.

Analysis of the GSR data obtained during each interrogation sequence was carried out to determine the name that yielded the highest conductance change within five seconds after a name presentation. A hit was scored whenever the conductance change to the actual name exceeded the maximum response to the alternative names in a given sequence.

The correct name detection rates ranged from .60 in the first repetition to .48 in the tenth repetition in the low-motivation group, from .55 to .50 in the high-motivation group, and from .60 in the first repetition to .42 in the tenth repetition in the high motivational countermanipulators' group.

Although there were fluctuations in the proportion of correct name identification, with repetitions, there was still a significant correct detection rate ($p < .05$) in the tenth trial that used a normal approximation to a binomial distribution (n, 1/5) in all groups. In addition to this analysis, the same decision rule previously mentioned was applied to the means of the response to the interrogation stimuli across repetitions, adding one repetition at a time. The proportions of correct name detection in the three experimental groups ranged from .60 with one repetition to .96 using all ten repetitions in the Lo group, from .55 to .93 in the Hi group and from .60 in the first repetition of the high motivational countermanipulators group to 1.00 using all ten repetitions.

DISCUSSION

The results of this experiment may be viewed as support for the research that has attempted to increase the efficiency of psychophysiological detection through a simple statistical combination of the reactivities obtained from several repetitions of the *same* interrogation sequence. To the best of the experimenters' knowledge, this is the first successful use of such an approach in the context of information detection. The reported finding has immediate applicatory implications: (*a*) it simplifies considerably the task of obtaining critical items for the information detection procedures in cases where these items are difficult to obtain, in addition, (*b*) it makes it easier to keep a few or one critical item from the knowledge of the population of potential suspects.

As to the fluctuations of the correct detection rates with repetitions, while there did not seem to be any particular trend from trial to trial in the low-motivation group, there appeared to be a tendency for the correct detection proportions to drop with further trials in the high-motivation group, and in particular, in the case of the high-motivation countermanipulators. This trend is indicated by a significant ($p < .05$) Spearman rank order correlation of —.82 between detection rate and repetition number for the high motivational countermanipulators group, —.55 for the high-motivation group, and a nonsignificant correlation of —.28 for the low-motivation group. However, a two-factor, Groups × Trials, repeated measures analysis of variance conducted on the conductance change data (ratio of the conductance change to the actual name, and the maximum conductance change to the alternative names) did not yield any significant effects. Thus, the possible interaction of the state of motivation and trials suggested by the differential correlations could not be detected using the individual data, and the continuous measure underlying the detection process. In addition, this analysis showed no differences in the actual conductance indices between the motivation groups.

The data of this experiment suggest that further consideration be given to the role of motivation in the detection of information. It is clear that independent measures of motivation are needed to evaluate the impact of instructions on the subjects. This study applied such a measure as exemplified in the identification of the high-motivational countermanipulations group. In contrast to Gustafson and Orne (1963) who reported insignificant detection under low motivation, the experimenters of this laboratory study have never observed insignificant detection rates except for certain groups of very young children (Lieblich, 1971), or uneducated Bedouin (Kugelmass & Lieblich, 1968). The present data did not indicate any difference in degree of detection that might have been expected to result from the specifically different instructions. Gustafson and Orne (1963) had used cards while this experiment used the subjects' names as critical items. Names have high relevance for a subject and are frequently used in his environment (for discussion of these facets see Lieblich, 1969, 1971). In cases where such stimuli are presented external motivation may not be critical. It is to be expected that in general, detection procedures would use highly relevant stimuli such as names, codes, etc. In such cases, it could be expected that the external motivational conditions would have limited effects.

REFERENCES

Ben Shakhar, G., Lieblich, I., & Kugelmass, S. Guilty knowledge technique: Applications of signal detection measures. *Journal of Applied Psychology,* 1970, **54**, 409–413.

Gustafson, L. A., & Orne, M. T. Effects of heightened motivation on the detection of deception. *Journal of Applied Psychology,* 1963, **47**, 408–411.

Kugelmass, S., & Lieblich, I. Relation between ethnic origin and GSR reactivity in psychophysiological detection. *Journal of Applied Psychology,* 1968, **52**, 158–162.

Lieblich, I. Identification of variables related to the efficiency of the detection of information within the organism using a physiological index. Unpublished doctoral dissertation, Hebrew University of Jerusalem, 1966.

Lieblich, I. Manipulation of contrast between differential GSR responses through the use of ordered tasks of information detection. *Psychophysiology,* 1969, **6**, 70–77.

Lieblich, I. Manipulation of contrast between differential GSRs in very young children. *Psychophysiology,* 1971, **7**, 436–441.

Lieblich, I., Kugelmass, S., & Ben Shakhar, G. Efficiency of GSR detection of information as a function of stimulus set size. *Psychophysiology,* 1970, **6**, 601–608.

Lykken, D. T. The GSR in the detection of guilt. *Journal of Applied Psychology,* 1959, **43**, 385–388.

Nevo, B. Mispar beayot beivkhun shkarim (Problems in the detection of deception). Unpublished master of arts thesis, Hebrew University of Jerusalem, 1968 (in Hebrew).

(Received October 10, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 116-118

RELATION OF PRICE TO PERCEPTION OF QUALITY OF NEW PRODUCTS

ARCH G. WOODSIDE[1]

University of South Carolina

Six groups of 12 construction workers were shown a new product, an electric lunch box, and provided with a printed advertisement that contained heating instructions and price information. Five different prices were used: $4, $5, $6, $8, and $10, with one group of subjects not receiving price information. The subjects individually inspected the product and completed (*a*) 11-point rating scales on advantage, use and quality and (*b*) a questionnaire selecting favorable and unfavorable words to describe the product. Results indicated that a linear regression model significantly fit the scaled responses for the treatments ($p < .01$).

Price has been indicated as a determining factor of perceived product quality whenever the consumer has little previous experience with the product or the potential risk and uncertainty involved in its use are great (Leavitt, 1954; Scitovsky, 1944–1945; Shapiro, 1968). Gabor and Granger (1966) have hypothesized a nonlinear price–perceived-product-quality relationship which has been supported by empirical findings by McConnell (1968) and Peterson (1970). However, two recent studies (Gardner, 1971; Rao, 1971) raise serious questions about any generalized price–quality relationship. These studies demonstrated that the salience of price in quality perceptions was negligible under conditions when other informational cues are available to consumers, for example, prior experience or consumer test reports.

It seems apparent that the value of price in consumer's perceptions of quality may depend on the availability of additional product information. The Stafford-Enis (1969) hypothesis would be a limiting case difficult to dispute: "It appears obvious to us that if price data provide the *only* clue to product quality, then perceived quality should be expected to vary directly with price [p. 456]."

The Gardner (1971) and Rao (1971) studies were performed with products that subjects could easily relate to prior purchase and use. Thus, the nonsignificant findings of price and perceived product quality might have been expected under such high-prior-information conditions.

Peterson (1970) used a product for which consumers had a relative lack of information: a recently developed but unmarketed soft drink concentrate packaged in a pressurized can which when mixed with water resulted in a novel carbonated soft drink (Gardner, 1971). Subjects in the experiment estimated quality in part by comparing the product with other soft drinks that they had purchased and used; however, direct prior experience with the product in the study was not possible. A significant price and perceived quality relationship was found by Peterson.

Thus, the relationship between price and perceived product quality may depend on the amount of prior information held by the consumer. The paradigm shown in Table 1 summarizes the argument and representative research findings.

The present study examines the relationship between price and perceived quality for a product for which consumers would have a relative lack of information. The selected product was hypothesized to be a "dynamically continuous" innovation (Robertson, 1971) about which consumers would not have direct prior knowledge but would be prone to compare its benefits to some perceived standard. Thus, a significant, nonlinear relationship was hypothesized.

METHOD

Product

The product used was a recently developed, though not yet marketed, electric lunch box that was made of reinforced plastic with a galvanized soleplate and a five-foot electric cord.

Subjects

The subjects were 72 construction workers living in the Columbia, South Carolina, metropolitan area. These subjects, who usually carried lunch boxes to work, had no prior knowledge of the product and were provided with little information as to its potential performance.

---

[1] Requests for reprints should be sent to Arch G. Woodside, College of Business Administration, University of South Carolina, Columbia, South Carolina 29208.

Procedure

The subjects were randomly assigned to one of six experimental conditions. Each subject was independently called aside during his lunch break and requested to participate in evaluating a new product. The product was displayed to the subjects, who were given a printed advertisement with the following headline: "hot box: electric, self-warming lunch box warms your lunch on the job!"

All information given to the subjects was contained on the advertisement, with price being the only variation. Five different prices were used: $4, $5, $6, $8, and $10. (The actual market price was to be $4.99 retail.) One group of 12 subjects served as a control group and did not receive any price information with their advertisement.

After inspecting the product, each subject completed a questionnaire by selecting 3 of 12 words "which best describe the electric self-warming lunch box to you." The following words were included in the questionnaire:

Needless
Strong
A fad
Great
Good idea
Worthless
Quality
Bad idea
Poorly made
Useful
Unreliable
Dependable

The subjects also completed 11-point rating scales on relative advantage, ease of use, and quality perceived in the product.

RESULTS

A total of 12 subjects were in each treatment condition. The subjects' word choice responses were separated into favorable and unfavorable selections. Favorable words included strong, great, good idea, quality, useful, and dependable. The percentages of favorable responses were 53, 78, 82, 89, and 90 for the $4, $5, $6, $8, and $10 treatments, respectively. A total of 71% of the word choices made by the control subjects were favorable, which is midway between the results of the price categories. This finding provides credibility to the control procedures, since one would expect that without knowledge of price, a neutral group would be expected to use the median number of favorable words used by the full range of groups, each of which was given a different price for the product.

The percentages of favorable and unfavorable responses were significantly different for the treatment groups using the chi-square test ($p <$ .01). The control group was not included in the analysis.

TABLE 1

REPRESENTATIVE RESEARCH FINDINGS ON THE RELATIONSHIP BETWEEN PRICE AND PERCEIVED PRODUCT QUALITY

Type of price-perceived-quality relationship	Amount and type of information		
	None	Comparative only	Prior experience
Significant, linear	Stafford and Enis (1969)		
Significant, nonlinear		Gabor and Granger (1966) Peterson (1970)	
Nonsignificant			Gardner (1971), Rao (1971)

The mean perceived quality ratings were 4.83, 5.33, 6.33, 8.67, and 7.33 for the $4 to $10 treatment groups, respectively. (The larger the mean, the greater the perceived quality.)

The calculated $F = 3.43$ ($df = 4/59$) for one-way analysis of variance was statistically significant ($p < .05$).

The Duncan multiple-range Test (Winer, 1962) was used to test for differences between individual treatment means. This analysis revealed that the mean quality ratings of subjects in the $4, $5, and $6 ranges were significantly less than the mean for the $8 treatment subjects ($p < .05$).

An orthogonal test (Winer, 1962) was performed to determine the appropriate form of the relationship. A linear regression model was found to fit the data significantly ($p < .01$), while a quadratic model was not significant.

DISCUSSION

The results partially support the hypothesis. A significant relationship was found between price and perceived product quality. The finding of a direct rather than a quadratic relationship may be due to limitations of the study, that is, an upper price limit may be near $10 for this product. Beyond the upper price limit, the product could be too expensive for quality perceived (Gabor & Granger, 1966). To test this hypothesis in the context of the present experiment, $12 and $14 price treatments would have been useful. Any upper limit on price might also be a function of the product's market, for example, executives or members of other higher income groups who carry their lunch to work may have a significantly higher price limit than the group of subjects studied.

The purpose of the present study has been to examine the conditions by which significant price–perceived-product-quality relationships would be found. The direct relationship found in the experiment is an extension of the Gabor and Granger (1966), McConnell (1968), and Peterson (1970) studies. Given a novel product to evaluate and a relative lack of information, the consumer may be expected to use price as an indication of quality. However, when additional information cues are available, the relationship between price and perceived quality may diminish. Within certain information limits, a generalized price–perceived-quality relationship may be specified.

REFERENCES

Gabor, A., & Granger, C. W. J. Price as an indication of quality: Report on an inquiry. *Economica,* 1966, **33**, 43–20.

Gardner, D. M. Is there a generalized price–quality relationship? *Journal of Marketing Research,* 1971, **8**, 241–243.

Leavitt, H. A note on some experimental findings about the meaning of price. *Journal of Business,* 1954, **27**, 205–210.

McConnell, J. D. The price–quality relationship in an experimental setting. *Journal of Marketing Research,* 1968, **5**, 300–303.

Peterson, R. A. The price-perceived quality relationship: Experimental evidence. *Journal of Marketing Research,* 1970, **7**, 525–528.

Rao, V. R. Salience of price in the perception of product quality: A multidimensional measurement approach. In, *Proceedings of the American Marketing Association.* Chicago: AMA, 1971.

Robertson, T. S. *Innovative behavior and communication.* New York: Holt, Rinehart & Winston, 1971.

Scitovsky, T. Some consequences of the habit of judging quality by price. *The Review of Economic Studies,* 1944–1945, **12**, 100–105.

Shapiro, B. P. The psychology of pricing. *Harvard Business Review,* 1968, **46**, 14–25.

Stafford, J. E., & Enis, B. M. The price–quality relationship: An extension. *Journal of Marketing Research,* 1969, **6**, 456–548.

Winer, B. J. *Statistical principles in experimental design.* New York: McGraw-Hill, 1962.

(Received August 23, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 119–121

A CONSUMER PROFILE ANALYSIS OF THE SELF-SERVICE GASOLINE CUSTOMER

J. M. McCLURG
Shell Oil of Canada

I. R. ANDREWS[1]
Simon Fraser University

Fifty drivers were interviewed at self-service gasoline stations and another fifty drivers were interviewed at nearby conventional service stations. Response differences between these two samples confirmed the predictions that the self-service customer would have an economic shopper orientation, that he would have the characteristics of people who are innovators in new product use, and that he would be less dogmatic in his thought processes. This last finding provided an interesting field replication of a multiproduct laboratory study by Jacoby.

Three hypotheses about possible differences between self-service and conventional service gasoline station customers were derived from a review of the general literature. The first hypothesis stated that because the self-service station features lower price and minimized contact with station personnel, the customers at these stations would be more like Stone's (1954) price sensitive economic shopper and less like his relationships oriented personalizing shopper. Since the self-service station was a recent innovation in the market area surveyed, it was further hypothesized that customers at these stations would tend to possess the characteristics associated with innovative product use. According to Rogers and Stanfield (1968) new product users tend to be better educated, more literate, have higher incomes and higher standards of living, have higher achievement aspirations for their children and are more knowledgeable about world and local events.

The third hypothesis was taken from a laboratory study by Jacoby (1971) who found that for many products, innovative use was inversely related to degree of dogmatism (Rokeach, 1960). On the basis of Jacoby's results and the fact that self-service stations were new in the area surveyed, the investigators of this study predicted that self-service customers would be less dogmatic in their thought processes than customers who patronize conventional gasoline stations.

METHOD

Fifty drivers were interviewed at five different self-service gasoline stations and 50 additional drivers were interviewed at five different conventional gasoline stations chosen because of their proximity to the self-service stations. The first part of the interview included four highly structured questions and one open-ended question about the customer's gasoline purchase practices. Also included were two semantic-differential type items, one about the kind of person "who would buy gasoline here" and one about "attitudes toward new products."

At self-service stations only there were additional questions which sampled customer opinion about such things as the intercom between the gasoline pumps and the cashier, knowledge about how much money was saved at a self-service station, etc.

The last part of the interview consisted of questions about age, amount of education, income level, etc. When the 15 to 30 minute interview was completed, the driver was given a questionnaire entitled, "Consumer opinions on social issues." This questionnaire included 40 items from Form-E of Rokeach's (1960) dogmatism scale. Since this scale includes a variety of exotic items such as, "It is better to be a dead hero than a live coward," the standard instructions were modified to include a consumer opinion phrase. Also the interviewer gave each subject a set of steel steak knives in appreciation of his participation in the study. Lastly, the prestamped return envelope was addressed to a department of economics and commerce rather than to an oil company or to a department of psychology. These efforts produced a very high return rate, 73 out of 100.

RESULTS

The economic shopper orientation of the self-service customer was verified by several pieces of data. For example, when the customers were asked in an open-ended question about why they decided to buy gas at the station where they were interviewed, 28 of the self-service patrons mentioned price while only 8 mentioned good service. By contrast, none of the conventional station patrons mentioned price and 21 mentioned good service.

---

[1] Requests for reprints should be sent to I. R. Andrews, Department of Economics and Commerce, Simon Fraser University, Burnaby, British Columbia, Canada.

When the self-service customers were asked to pick the 5 most important reasons out of a list of 12 possible reasons for buying gas at a particular service station, 23 of the self-service patrons assigned the number one rank to price and 11 more gave price a rank of two. Among the conventional station customers, pleasant staff received 16 first place rankings and 9 second place rankings. Two additional indications of the economic shopper orientation of the self-service customers were their much greater tendency to purchase gas at other kinds of discount stations and the fact that 48 out of 50 self-service patrons knew that they were buying a premium quality product at some discount in price.

The hypothesized tendency for greater product innovativeness among the self-service customers was examined with two semantic-differential type questions and with demographic data. The first question asked the respondent to imagine the kind of person who would be apt to buy gas at the particular station where the interview was conducted. On four out of six semantic-differential scales, the group means for self-service and conventional service customers were significantly different at the .05 level ($df = 98$, one-tailed). In their projected self-image, the self-service customers imagined themselves as younger, bolder, more lively, and more modern.

The second semantic-differential type question —attitudes toward new products—produced significant group mean differences on five out of seven scales ($df = 98$, one-tailed). The self-service patrons were more apt to be early adopters, were more interested in new products, more carefree, more influential, and more excited by new products.

When the two samples were compared on demographic items the conventional service customers were more apt to own two cars and to drive a luxury car or a sports car, and they had a slightly higher income. The self-service patrons were younger, more apt to be students and somewhat better educated. The tendency for greater education in the innovative group was consistent with data reported by Rogers and Stanfield (1968) but the difference in income level and standard of living was not. Presumably this departure from previous results was produced by the cost savings associated with the new product sampled in the present study.

To test the third hypothesis a separate dogmatism score was calculated for each of the 73 subjects who completed and returned the "Consumer opinions on social issues" questionnaire (in reality, Form-E of Rokeach's dogmatism scale). As predicted the mean score for 41 self-service customers (−15.85) indicated a group tendency for low dogmatism. By contrast, the mean score for 32 conventional service respondents (−3.37) was quite near the zero midpoint of the dogmatism scale. Included among the 32 conventional service respondents were six persons who reported making only 20% to 30% of their gas purchases at conventional stations whereas the modal members of that group made 80% to 90% of their purchases at conventional stations. When these atypical gas purchasers were deleted, the mean dogmatism score for conventional service respondents was plus .67, almost exactly on the scale midpoint. Without the deletion of atypical gas purchasers from the conventional service group, the mean difference between groups was not significant ($t = 1.58$, $df = 71$). With their deletion, however, the difference in mean dogmatism scores for the two groups was significant at the .05 level ($t = 2.03$, $df = 65$, two-tailed).

An alternate nonparametric analysis with all 73 questionnaire respondents showed that the proportion of subjects with dogmatism scores above the scale midpoint (zero) was quite different for the two groups—17 out of 32 for the conventional service customers versus only 12 out of 41 for the self-service customers. This difference in proportions was highly significant ($Z = 2.08$, $p < .05$).

Also confirming the personality difference between the two samples was the difference in the proportion of subjects who returned the dogmatism questionnaire—32 out of 50 conventional station interviewees versus 41 out of 50 self-service interviewees ($Z = 2.02$, $p < .05$).

DISCUSSION

As predicted from the cost-saving and depersonalized nature of the self-service operation, the drivers who were sampled at self-service gasoline stations had a strong economic shopper orientation. Also, as predicted from the fact that self-service stations were still new in the region surveyed, the self-service drivers had several of the characteristics of people who are innovators in new product use. Lastly, the results provided field research confirmation of Jacoby's (1971) laboratory finding that new product users are less dogmatic in their thought processes than persons who prefer established products.

REFERENCES

JACOBY, J. Multiple-indicant approach for studying new product adopters. *Journal of Applied Psychology*, 1971, **60**, 384–387.

ROGERS, [illegible] M., & STANFIELD, J. D. Adoption and diffusion of new products: Emerging generalizations and hypotheses. In F. M. Bass, C. W. King & E. Pessemier (Eds.), *Application of the Sciences in Marketing Management*, New York: Wiley, 1968.

ROKEACH, M. *The open and closed mind.* New York: Basic Books, 1960.

STONE, G. P. City shoppers and urban identification: Observations on the social psychology of city life. *American Journal of Sociology*, 1954, **60**, 36–45.

(Received October 10, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 122–124

MOTIVATIONS OF BLOOD DONORS AND NONDONORS[1]

ROBERT M. OSWALT[2] AND MARIE NAPOLIELLO

Skidmore College

A survey of over 85% questionnaire returns from a random sample of two college populations indicated that 40% were blood donors (compared with 3% nationally) and 60% nondonors. Motivations for donors were: humanitarian, peer pressure, and curiosity experience. Motivations of nondonors were: legitimate medical excuses, fear of needle, general apprehension, and fear of aftereffects. Recommendations for blood procurement agencies are discussed.

Population growth, advances in surgery and medicine, and expansion of hospital facilities have significantly increased the demand for whole human blood (American National Red Cross, 1972). As a result of this increased blood usage, the American Association of Blood Banks ("The Blood Business," 1972) has estimated that 8.5 million units of blood are needed every year. Nonetheless, only 3% of the eligible population is now donating blood, whereas 4% is required to meet the need (American National Red Cross, 1972). Often the supply of blood lags behind the demand because the source depends entirely on the voluntary donation of individuals.

Although there is an increasing need for blood, information concerning the motivation of blood donors is scarce. Most prior studies have investigated social, economic and demographic characteristics of donors (Grace, 1957; Korzekwa, Jordan, & Alsever, 1960; London & Hemphill, 1965; Phillips, 1961), either compared the characteristics of first-time versus repeating donors (Condie & Maxwell, 1970a), or studied the differences between voluntary and paid donors (Condie & Maxwell, 1970b). In general, the average donor is male, younger and of a higher socioeconomic and educational level than the average population, gives humanitarian and personal reasons for donating, and expresses little or no concern toward negative aspects of donating.

Knowledge concerning the motivations that inhibit the donation of blood is a largely unexplored area. Even though 97% of the population of the United States does not volunteer to donate blood (American National Red Cross, 1972), no study has ever been conducted in the United States on the motivation of *nondonors.*

The purpose of this study was to examine motivations of a random sample of blood donors *and* nondonors. All of the previously mentioned studies limited their subjects to those who made a previous commitment to donate blood. Information about why people do or do not donate blood could be of practical significance in the procurement of new donors and in planning future educational and motivational programs.

METHOD

Subjects

Ten percent of the entire populations of two eastern liberal arts colleges were chosen for this survey by taking every tenth name in their student directories, beginning at a randomly chosen name. The sample consisted of 187 females and 152 males.

Procedure

A questionnaire and return-addressed, stamped envelope was sent to each subject with a cover letter explaining the purpose of the survey and emphasizing its significance. Subjects were assured that all responses would be treated confidentially.

The questionnaire requested demographic data as well as specific data on the subject's age, whether or not the subject ever donated blood, geographic location of first blood donation, reasons for donating, the total number of times donated, the presence of any negative feelings before and after donating, whether the subject donated for remuneration or not, and self-opinions concerning the effect of incentives.

RESULTS

Eighty-seven percent of the subjects in the sample returned usable questionnaires and their responses indicated that approximately 40% of the sample had donated blood while 60% had not. The fact that 40% of the subjects were donors differs significantly from the American National Red Cross estimate that only 3% of the eligible national population are donors.

---

[1] This study was supported in part by National Science Foundation Grant GY-9965. The authors wish to thank Robert P. Mahoney for his contribution to initial discussions on the problems associated with obtaining blood donors.

[2] Requests for reprints should be sent to Robert M. Oswalt, Department of Psychology, Skidmore College, Saratoga Springs, New York 12866.

Most donors in this survey (39%) have donated only one time (although one third of the donors reported they would donate as often as every three months), have donated at Red Cross blood drives at college (89%), and have donated for no remuneration (97%).

Donors reported the following as major reasons for donating blood: humanitarian reasons (66%), peer pressure (19%), convenience of college bloodmobile (18%), and curiosity in experiencing the process of donating blood (13%). Prior to donating they reported they were nervous (77%), afraid of the needle (37%), and concerned about feeling weak, faint, or dizzy (28%). Subsequent to donating they were less fearful, but essentially afraid of the same aspects—nervousness (23%), needle (24%), and faintness (11%).

The majority of the nondonors (60%) reported that legitimate medical reasons (underweight, hepatitis, mononucleosis, low blood pressure, and the use of medication) prohibited them from donating blood. Other reasons which deterred nondonors were general nervousness and fear (37%), fear of the needle (27%), inconvenience (19%), and apathy (16%).

Eighty percent of the donors and nondonors reported that incentives (e.g., money, free replacement blood if needed) would not increase or decrease *their* rate of giving, but cash payments or income tax deductions would be the best incentives to encourage other people to donate. The Federal Government's present policy (*Federal Tax Guide,* 1973) which requires one to pay tax on money received for blood sold (property) and does not allow one to deduct the value of blood donated free (service rather than property) was seen as unreasonable by the majority of respondents.

DISCUSSION

The results of this study suggest the following recommendations to increase the donation of blood:

1. Blood procurement agencies should contact more college students as perspective donors in order to satisfy the demand for blood. This study indicated that approximately 40% of the sample college populations donated blood compared to only 3% of the eligible national population. If 40% of the approximately eight million college students (*Chronicle of Higher Education,* 1972) in the United States were to give only one pint of blood a year, over three million pints could be obtained. This one source (colleges) could provide almost the entire yearly requirements of the Red Cross and one third of the national needs. Yet 1971 figures indicate that the Red Cross obtains less than 8% of its blood supply from colleges and universities (American National Red Cross, 1971).

2. Blood procurement agencies should make the donation of blood more convenient by locating bloodmobile units on college campuses. Since 89% of the donors reported that their first donations were at college bloodmobile units, the convenient placement of bloodmobiles could influence more individuals to donate blood. Very few students went to hospitals or bloodmobile centers to donate blood.

3. Blood procurement agencies probably could obtain more donors if they made an effort to contact donors at some specified intervals. This survey indicated that 34% of the donors would donate blood regularly every three months *if called upon,* and others would donate at longer intervals *if contacted.* Perhaps cards could be mailed at intervals agreed upon by the donors to remind them that they are eligible to give blood again. These cards could be patterned after the cards now sent to remind an individual that he is due for a dental appointment. In addition, individual appointments could be arranged in order to eliminate waiting and make the specified time more of an obligation.

4. Blood procurement agencies might consider eliminating the emphasis on emergency situations as a means of motivation and begin to use financial incentives to encourage donations. Emergencies, as incentive factors, received low percentages from both donors (10%) and nondonors (15%). In fact, 80% of the respondents reported that the use of *any* incentive would not change their rate of donation. However, the other 20% could be motivated to donate if monetary rewards were given either in cash or as income tax deductions.

These suggestions are, of course, based on the results of the two colleges sampled for this survey. Whether or not these results would generalize to other colleges and universities is an empirical question which could and should be answered.

REFERENCES

AMERICAN NATIONAL RED CROSS. Blood center operations, June 30, 1971.

AMERICAN NATIONAL RED CROSS. Red Cross blood program, 1972.

The blood business. *Time*, September 11, 1972, pp. 47–48.

CHRONICLE OF HIGHER EDUCATION, April 17, 1972, p. 1.

CONDIE, S., & MAXWELL, N. Social psychology of blood donors. *Transfusion*, 1970, **10**, 79–83. (a)

CONDIE, S., & MAXWELL, N. Comparative demographic profiles: Voluntary and paid blood donors. *Transfusion*, 1970, **10**, 84–88. (b)

FEDERAL TAX GUIDE. Englewood Cliffs, N.J.: Prentice-Hall, 1973, 16,015 (Revenue Ruling 162, 1953-2, CB 127).

GRACE, H. A. Blood donor recruitment: A case study in the psychology of communication. *Journal of Social Psychology*, 1957, **46**, 269–276.

KORZEKWA, F. W., JORDAN, W. Q., & ALSEVER, J. B. The blood donor: Who are our blood donors? An analysis of social and other characteristics of 12,759 donors. *American Journal of the Medical Sciences*, 1960, **240**, 36–47.

LONDON, P., & HEMPHILL, B. M. The motivations of blood donors. *Transfusion*, 1965, **5**, 559–568.

PHILLIPS, E. A study in blood donor motivation. *The Medical Journal of Australia*, 1961, **48**, 742–748.

(Received September 18, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 1, 125–126

SOME FOLKWAYS OF POLITICAL LEADERS: THE RULES OF THE GAME

ROBERT N. SPADARO[1]

Temple University

This study compares opinions about political folkways of political leaders and of voters before, during, and after an urban mayoralty election. Twenty politicians and 141 voters were administered a questionnaire composed of eight specific folkways. The major findings are (*a*) politicians and voters were substantially in agreement on seven out of the eight folkways, and (*b*) voters increasingly, over the three time periods, rejected the effectiveness of personal contact, while politicians remained strongly in support of such campaigning. It is, thus, suggested that politicians adhere to specific campaign folkways regardless of voter preference. The significance of this trend in understanding political campaigning is discussed.

What are the rules, or folkways, used by politicians as cues to social and policy positions and manipulative strategies? While we know a good deal about voter behavior and experience, we have little systematic knowledge of the campaign experience for the politician; thus, our understanding of this important voter–politician linkage is limited.

Recent studies have attempted to describe folkways of national and state legislators (Matthews, 1960; Wahlke, Eulau, Buchanan, & Ferguson, 1962). Although there have been fewer systematic attempts to examine such folkways for the politician campaigning for office, the work of Levin (1962), Eldersveld (1966), and others have indicated that such folkways do exist and that they provide constraints on the behavior of politicians. Perhaps these constraints tend to preclude learning experiences during the campaign, and politicians' performance of such folkway behaviors may lend a ritualistic quality to campaigns. The present exploratory study (*a*) identifies some specific campaign folkways, (*b*) examines the extent to which the opinion of politicians regarding specific political folkways agrees with that of voters, and (*c*) determines possible changes in these opinions during a campaign.

METHOD

One hundred and eighty politicians from both major parties, active in local, state, and national politics, were requested to give political "sayings" they believed were applicable as the "do's and don'ts" of political campaigning in four contrasting states (California, Minnesota, Pennsylvania, and South Carolina). Their most frequent responses were then cross-tabulated into eight specific campaign folkways and resubmitted to the original sample who unanimously (170 answered) agreed that these were basic campaign rules. The second stage involved the presentation of these sayings, by questionnaire, to 20 politicians and 208 voters; of these, 141 completed all three interviews before, during, and after a Philadelphia mayoralty election. Politicians were asked, "Do these following statements guide your behavior?" Voters were asked, "Do the following sayings or actions by politicians attract your vote?" Controls for socioeconomic status differences (income, education, occupation, race, and ethnic background) were applied to voters' perceptions. Differences between samples were measured by chi-square.

RESULTS

The results are summarized in Table 1. The primary difference found between politicians and voters was on the folkway pertaining to effectiveness of personal contact. Differences between samples at each stage of the campaign differed significantly (.05 level) on this issue. In contrast to previous findings in the social sciences (e.g., Cutright & Rossi, 1958), limited to contacts by committeemen, voters here rejected the rule pertaining to the effectiveness of personal contact—"press the flesh" campaigning. Table 1 shows that prior to the campaign, voters marginally supported this rule (53%), but over time a majority of the voters increasingly changed their opinion (48% and 40%), while the politicians were in higher agreement initially and remained relatively constant in their support at all three stages. Table 1 also suggests that politicians increasingly

---

[1] Requests for reprints should be sent to Robert N. Spadaro, Department of Political Science, Temple University, 1949 Broad Street, Philadelphia, Pennsylvania 19122.

TABLE 1

POLITICIANS' AND VOTERS' AGREEMENT WITH CAMPAIGN FOLKWAYS AT THREE STAGES

Folkway	Politicians (% who agree)			Voters (% who agree)		
	Campaign stage			Campaign stage		
	1	2	3	1	2	3
A candidate has got to get out and meet the people—shaking hands, supermarkets, walking tours, and so forth.*	80	70	80	53	48	40
People vote their pocketbook.	70	80	80	91	92	97
A scandal, exposé, in short a bombshell during the campaign against your opponent or his party is highly effective.	90	100	80	89	89	88
The majority of voters usually have made up their minds before the campaign even begins, so you have to concentrate on the swing vote.	60	80	60	58	42	68
You can't beat somebody with nobody.	80	80	70	94	90	94
Stick to one or two issues in campaigning instead of several.	70	80	100	83	79	79
People vote against not for.	70	80	70	63	71	77
You have to peak at the right time.	70	80	70	66	71	69

* $p < .05$ between samples for each of the three stages.

agreed that "it's better to stick to one or two issues." Voters also tended to agree more at all stages that "you can't beat somebody with nobody," and "people vote their pocketbook." There were no differences on any folkway between voters on the basis of party, income, or ethnic–racial variables. The fact that more differences between politicians and voters were not found indicates some general saliency to these folkways and to the politician's insights on general campaign strategies.

The one issue on which there was substantial change in the opinions of the politicians was in the need to stick to one or two issues; politicians increasingly agreed with this folkway as the campaign proceeded. In general, however, the campaign resulted in little change in politicians' agreement with these folkways.

These results are at least consistent with earlier findings (McClosky, Hoffman, & O'Hara, 1960) that politicians are groups of like-minded men. Smith (1958) found that when politicians were questioned as to whom they most communicated with during a campaign for information, they cited fellow politicians as their first source, with their secondary source being family and friends. This social reinforcement from colleagues may form the primary feedback linkages for fellow politicians (Dexter, 1963), with resulting constraints on changes due to politician–voter interactions during a campaign.

REFERENCES

CUTRIGHT, P., & ROSSI, P. H. Grass roots politicians and the vote. *American Sociological Review,* 1958, 23, 171–179.

DEXTER, L. A. The representative and his district. In W. Polsby, R. A. Dentler, & P. A. Smith (Eds.), *Politics and social life.* Boston: Houghton Mifflin, 1963.

ELDERSVELD, S. J. *Political parties: A behavioral analysis.* Chicago: Rand McNally, 1966.

LEVIN, M. B. *The complete politician.* Indianapolis: Bobbs-Merrill, 1962.

MATTHEWS, D. R. *U. S. Senators and their world.* Chapel Hill: University of North Carolina Press, 1960.

MCCLOSKEY, H., HOFFMAN, P. A., and O'HARA, R. Issue conflict and consensus among party leaders and followers. *American Political Science Review,* 1960, **54**, 406–427.

SMITH, M. B. Opinions, personality and political behavior. *American Political Science Review,* 1958, **52**, 210–215.

WAHLKE, J. C., EULAU, H., BUCHANAN, W., & FERGUSON, L. C. *The legislative system: Explorations in legislative behavior.* New York: Wiley, 1962.

(Received May 22, 1972)

JOURNAL OF APPLIED PSYCHOLOGY

Copyright © 1974 by the American Psychological Association, Inc.

April 1974 Vol. 59, No. 2

ARTICLES

Effects of Member Participation and Commitment in Group Decision Making on Influence, Satisfaction, and Decision Riskiness *Michael R. Cooper and Michael T. Wood* 127

Brainstorming Procedure, Group Size, and Sex as Determinants of the Problem-Solving Effectiveness of Groups and Individuals *Thomas J. Bouchard, Jr., Jean Barsaloux, and Gail Drauden* 135

A Causal–Correlational Analysis of the Job Satisfaction and Performance Relationship *John P. Wanous* 139

Sources of Satisfaction and Dissatisfaction among Solid Waste Management Employees *Edwin A. Locke and Roman J. Whiting* 145

Effects of Applicant Stereotypes, Order, and Information on Interview Impressions *Manuel London and Milton D. Hakel* 157

Comparability of Interviewer, Group, and Individual Interview Ratings *Larry F. Moore and Alec J. Lee* 163

Participation Effects in Appraisal Interviews Conducted in a Training Situation *Joseph M. Hillery and Kenneth N. Wexley* 168

Extensions of a Path–Goal Theory of Motivation *Martin G. Evans* 172

Race, Economic Class, and the Intention to Work: Some Normative and Attitudinal Correlates *Jack Feldman* 179

Improving Job Performance Through Training in Goal Setting *Gary P. Latham and Sydney B. Kinne III* 187

Aspects of Driving Performance as a Function of Field Dependence *Paul L. Olson* 192

A View of What Can Happen When Behavioral Expectation Scales Are Developed in One Setting and Used in Another *Walter C. Borman and W. Robert Vallon* 197

What Does the Least Preferred Co-Worker Scale Really Measure? A Cognitive Interpretation *Martin G. Evans and Jerry Dermer* 202

Predicting Managerial Success of Master of Business Administration (MBA) Graduates *Alan G. Weinstein and V. Srinivasan* 207

Pupillography in Industrial Fatigue Assessment *Tamara Geacintov and W. Scott Peavler* 213

SHORT NOTES

Getting the Mail Questionnaire Returned: Some Practical Research Considerations *John F. Veiga* 217

Effects of Alternative Follow-Up Procedures on Mail Survey Response Rates *Michael J. Etzel and Bruce J. Walker* 219

806(3)
18/9/74
Lib

Using Monetary Inducements to Increase Response Rates from Mailed Surveys: A Replication and Extension of Previous Research
Schuyler W. Huck and Edwin M. Gleason 222

A Comparison of Individual, Subgroup, and Total Group Methods of Problem Solving
Thomas J. Bouchard, Jr., Gail Drauden, and Jean Barsaloux 226

Expectancy as a Moderator of the Relationship between Job Attitudes and Turnover
Fred Dansereau, Jr., James Cashman, and George Graen 228

Job Expectations and Valences of Job Rewards for Culturally Disadvantaged and Advantaged Clerical Employees *Richard D. Arvey and Stephen J. Mussio* 230

Development of a Scale to Assess Student Nurses' Views of the Professional Nursing Role
L. M. Crocker and B. J. Brodie 233

Structure and Measurement of Marine Recruit Attitudes
Richard F. Booth and Anne Hoiberg 236

A Disguised Measure of Civil Rights Attitudes *Abraham K. Korman* 239

Self-Image and Stereotypes of Femininity: Their Relationship to Women's Role Conflicts and Coping
Francine E. Gordon and Douglas T. Hall 241

Effects of Expectation, Social Class, and Experience on In-Service Teacher Behavior in Small Groups
Len Sperry 244

Effect of Terminal Ranking Error on Meaningful Occupational Choice
Edwin E. Wagner and Thomas O. Hoover 247

Development of Behaviorally Anchored Rating Scales as a Function of Organizational Level *Sheldon Zedeck, Nicholas Imparato, Moshe Krausz, and Timothy Oleno* 249

LIST OF MANUSCRIPTS ACCEPTED 154, 144

ERRATUM 252

INSTRUCTIONS TO AUTHORS

Articles must possess: (*a*) conciseness; (*b*) unambiguous and simple vocabulary; (*c*) conformity to accepted technical style in tables, terminology, and references; and (*d*) conclusions that are clearly related to the evidence presented. Authors must progress from a statement of problem or purpose, through analysis of evidence, to conclusions and implications.

Criteria for evaluation include: (*a*) significance of contribution, (*b*) technical adequacy, (*c*) appropriateness for the *Journal of Applied Psychology*, and (*d*) clarity of presentation.

See the *Publication Manual of the American Psychological Association* (1967 Revision), $1.50, for the format of manuscripts. Articles not prepared in this manner cannot be reviewed. Note sections on typing (p. 48, *all copy must be double-spaced*), tables, figures, and references. The first author's last name must be typed on the upper left-hand corner of each page, except for blind reviews. Do not abbreviate, for example, journal titles in reference lists, the word "subject," names of groups, etc. Each manuscript copy must include an abstract of 100–120 words typed on a separate sheet of paper.

Optional blind review must be specifically requested when a manuscript is first submitted. The author's name and affiliation must appear only on a separate title page, included with each copy of the manuscript. Footnotes containing the identity of the author or his affiliation must be on separate pages.

Check recent issues of the *Journal of Applied Psychology* for approximate length of regular articles. (One printed page equals three double-spaced manuscript pages.) A few longer articles of special significance are printed as monographs. Short Notes feature brief reports on studies such as those involving some methodological contribution or important replication.

The original drawing of a figure or an 8 × 10 inch glossy print must be submitted with the original copy of the manuscript. Duplicate copies may be photographic or pencil drawn. Figures *must* be hand-lettered.

Supplementary materials formerly deposited with the National Auxiliary Publications Service need no longer be submitted with the manuscript. Retain supporting and raw data for at least 5 years after publication of the article.

Journal of Applied Psychology
1974, Vol. 59, No. 2, 127–134

EFFECTS OF MEMBER PARTICIPATION AND COMMITMENT IN GROUP DECISION MAKING ON INFLUENCE, SATISFACTION, AND DECISION RISKINESS[1]

MICHAEL R. COOPER[2]
GTE Laboratories, Inc., Waltham, Massachusetts

MICHAEL T. WOOD[3]
Ohio State University

Forty laboratory groups, each consisting of three members, were arranged to discuss structured decision issues. Participation in only one of three decision-making phases (generation, evaluation, or choice) constituted partial participation, in contrast to complete participation. Group members were committed or not committed to carrying out the decisions reached. Results indicated that perceived intragroup influence and satisfaction were greatest with complete participation. With partial participation, influence and satisfaction were greatest in the choice phase. Although perceived influence was greater under commitment conditions, satisfaction was greater under noncommitment conditions. Decision riskiness was affected by neither treatment.

Research and theory in social and organizational psychology have suggested that participation in group decision making enhances members' influence and satisfaction. Further, group decision making often elicits riskier courses of action than does individual decision making, although individuals who are more committed to their decisions tend to make more cautious decisions. These general formulations provided a base for the specific hypotheses tested in this study.

Group decision making, which has been a widely advocated means of sharing power (Likert, 1967; Tannenbaum, 1968; Wood, 1973), has recently been conceptualized and studied as a multiphased process of varying activities in which group members may participate. Based on theoretical descriptions of organizational decision making and empirical assessments of group problem solving, Wood (1972a) formulated a three-phased process involving the generation, evaluation, and choice of alternative resolutions to specific decision issues. Wood's research showed that the influence of members and the total group (leader and members) influence pie were greater when members participated throughout the process than when they participated in only two phases. Results suggested further that members' influence was weakest when they were unable to participate in the choice phase.

In the present study, participation was manipulated within the context of the three-phase process. However, partial participation conditions were designed to permit member participation in only one phase, thus reflecting commonly cited organizational participation programs, for example, suggestion boxes. The participation conditions were defined as follows: (*a*) generation, evaluation and choice—complete participation with group members able to participate in each of the three phases; (*b*) generation—members participate only in generating alternatives (a theoretical external group, analogous to upper-level organizational decision-making groups, was responsible for the evaluation and final choice of alternatives partially based on the alternatives generated by the experimental group); (*c*) evaluation—members participate only in evaluating alternatives, with the external group being responsible for generating alternatives and later making final choices partially based on the experimental group's evaluations; and (*d*) choice—members participate in choosing alternatives based on the external group's generation and evaluation.

---

[1] This study is based on a portion of the first author's dissertation submitted as partial fulfillment of the requirements for the PhD to the Ohio State University (Cooper, 1972). The contributions of Milton D. Hakel, S. Rains Wallace, and Robert J. Wherry, Sr., are gratefully acknowledged.

[2] Requests for reprints should be sent to Michael R. Cooper, GTE Laboratories, Inc., 40 Sylvan Road, Waltham, Massachusetts 02154.

[3] Now at Battelle, Human Affairs Research Centers, Seattle, Washington.

The following hypotheses concerning the effects of these participation variations on members' influence, satisfaction, and the riskiness of decisions were tested.

1. Perceived member influence and satisfaction are greatest with complete participation. When members have participated in only one phase of decision making, their influence and satisfaction are greatest when that phase involves final choice of alternatives.

2. The riskiness of group decisions is greatest when members have participated throughout the process. Decision riskiness does not vary across forms of single-phase participation. Previous research suggests that increased group interaction permits an increased diffusion of responsibility leading to a shift toward more risky decisions (Pruitt, 1971). Since the generation, evaluation, and choice condition implies a greater interaction potential, compared to single-phase participation, riskiness is greatest in the generation, evaluation, and choice condition.

Commitment to decisions has, for the most part, been investigated as a dependent variable. The studies that have used commitment as an independent variable have suggested that people produce more cautious decisions when they anticipate personal commitment to their choices (Dion, Miller, & Magnan, 1971; Mann, 1971). Similarly, seriousness of decision consequences has been found negatively related to shifts toward risk, particularly when there are negative consequences associated with risky choices (Clark & Willems, 1969; Johnson & Andrews, 1971). The following hypotheses were tested concerning effects of decision commitment on influence, satisfaction, and risk taking.

3. Perceived member influence and satisfaction are greater when group members are committed to carry out the actions dictated by their decisions than when they are not so committed.

4. Group decision riskiness is greater when members are not committed to carry out the actions dictated by their decisions than when they are committed.

METHOD

Overview

Effects of participation and commitment in decision making were investigated in small laboratory groups. Forty groups, with three members each, discussed structured decision issues under conditions of varying participation (generation, evaluation, and choice; generation; evaluation; or choice conditions) and commitment (committed or noncommitted). Questionnaires were developed and administered to measure pretask perceptions of task importance and posttask perceptions of participation, commitment, influence, and satisfaction. The principal hypotheses were tested with 4 × 2 analyses of variance in questionnaire responses.

Participants

Group members were 120 general psychology students at Ohio State University who participated for required research credit. There were 67 females and 53 males. Their mean age was 18.9 years; 94% were freshmen or sophomores. Of the 40 groups, 32 were sexually heterogeneous, 5 were all female, and 3 were all male. Seventy-eight percent of the people did not know either of their comembers prior to the experiment. Participants were randomly assigned to treatment conditions, except that those participating for two hours instead of one hour were assigned to complete participation. There were five group replications per condition.

Decision Task

The task focused on determining attitude statements to be included in a campus survey of student attitudes. Group members were instructed to select, evaluate, and/or make final choices of statements for inclusion in the survey. They were also told that they would conduct the survey themselves in face-to-face interviews with students on campus (commitment conditions) or that other students would conduct the survey (noncommitment). This task was intended to represent organizational decision tasks and to be relevant to student participants.

Determination of attitude topics and items to comprise the task was accomplished by having 260 undergraduate students generate four topics they felt should be posed to students in an attitude survey and also having them generate 6 items designed to tap attitudes within each topic. They were asked to develop two "risky" items, two low-risk items, and two intermediate-risk items per topic. This procedure yielded a pool of topics and items that was reduced to six topics with 12 items each. These 72 items were then presented in random order to a sample of 15 industrial-organizational psychology graduate students who scaled the items on a risk dimension. They were asked to evaluate the potential riskiness of each item by indicating the probability, on an 11-point scale, of a college student interviewer *not* wanting to personally confront another college student with the attitude question. This definition of risk is similar to that of Belovicz and Finch (1971) who view risk as a function of the desirability and probability of decision consequences. The graduate students similarly evaluated the riskiness of the six topics without reference to specific items.[4]

The final task consisted of items measuring attitudes toward sex (topic risk value = 6.2)[5], abortions (4.9), racism and discrimination (4.8), drugs (4.7), the military draft and Vietnam (2.1), and ecology (1.0). There were 12 items per topic, and the items varied along the risk continuum. For example, high-, intermediate-, and low-risk questions for the topic of sex included (*a*) "Have you ever had sexual relations with a member of the same sex as your own?" (risk value = 9.6); (*b*) "Do you believe that it is important to marry a virgin?" (4.3); (*c*) "Should sex-education courses be offered in the public schools?" (.7).

Pretask and Dependent Measures

Prior to the task, members completed a background form including measures of participant characteristics and task importance, that is, "Do you believe that a major survey of student attitudes is or can be important and should be conducted? (yes or no)?"

After the decision-making session, members completed a feedback form, consisting of (*a*) treatment manipulation checks, (*b*) perceived influence, and (*c*) satisfaction. The manipulation checks consisted of the following items to which members responded on 9-point scales ranging from "completely true" to "not at all true": (*a*) participation in the generation condition—"I was able to participate in the preliminary selection of items"; (*b*) participation in the evaluation condition—"I was able to actively discuss my opinions, choices, etc., with the other members of my group"; (*c*) participation in the choice condition—"I was able to play a direct part in the final choice of questions to be actually asked of students on campus"; (*d*) perceived commitment—"I realize that I will actually be conducting interviews with students on campus." The influence measure asked members to distribute a constant sum of (100) hypothetical influence points among the three members of his group and any other decision-making groups. Satisfaction was measured by 12 items with 9-point scales (completely satisfied-completely dissatisfied, with reversed anchors for half of the items). Items concerned the decision outcomes, one's own role in the group, the decision-making procedure, intermember relationships, and the decision task. Decision risk was determined on the basis of the mean scale values of attitude items included in the groups' selections, evaluations, or choices.

Treatment Manipulations

Treatments were manipulated through instructions administered by advanced undergraduate research assistants. They provided members with an orientation to the task and rationale for their involvement in the attitude survey.[6] For commitment conditions, they were told that they would be the actual students asking questions of their fellow students on campus, in face-to-face individual interviews. The members in noncommitment conditions were told that the survey itself would be conducted with *other students, not themselves*, doing the actual interviewing. Members were debriefed after the experiment, and no interviews were actually conducted.

Instructions were also given for each participation condition. For example, the basic manipulation portion of instructions for the generation groups was as follows:

> Here is a list of six topics with 12 questions on each, which we've generated to provide guidelines for your selection (7) of questions during the limited amount of time available. During the next 20 minutes, we'd like you to select three topics with only six questions for each topic. Your selections should be those questions you feel are most relevant to the survey, even though these represent merely preliminary selections and not final choices. *Another group* (8) will evaluate your selections and decide on a final set of two topics with three questions for each topic, which will be in-

---

[4] Items were also scaled for quality on a 5-point scale of comparative goodness–poorness as an attitude item. A curvilinear risk-quality relationship was initially expected with high-quality items being of intermediate risk. However, in this scaling sample, risk and quality were strongly, inversely, and linearly related ($r = -.51$, $p < .01$). Moreover, treatment effects were essentially opposite for decision riskiness and quality. Thus, decision quality is not dealt with as an independent criterion in this report.

[5] Risk values based on topic evaluations and on the mean item-risk values for a given topic were equivalent.

[6] Detailed instructions may be obtained from the authors.

[7] Because of the need for objective measurement of decision risk across all conditions, the topic and item domain were specified a priori, and resulted in a modified "structured generation of alternatives" decision phase or selection.

[8] The other group did not exist. The experimenter served as coordinator throughout group sessions creating artificial feedback, which the other group supposedly produced and which was transmitted to members by the research assistants. This group was created to function as a perceived source of varying amounts of influence, in the manner of upper level decision-making groups that remain external to most organizational members yet usually exert great influence in organizational decision making.

cluded in the final survey. This other group is made up of graduate students who are more qualified and trained in evaluating attitude items and making final choices. However, your input is important because the other group will evaluate and choose from those topics and items which you select.

Instructions were similarly developed for the evaluation and choice groups. In each case, the groups spent 20 minutes on the task. During the time that the hypothetical other group was processing their information, the members individually performed a filler task involving the specification of campus locations for the survey interview. The generation groups received feedback on the other group's choices of two topics with three items each. The evaluation groups were provided with Three Topics × Six Questions presumably generated by the other group and (later) the other group's choices based on their evaluations. They evaluated the items on 4-point scales of relevance, importance, and informativeness.[9] These groups began the evaluation immediately and performed the filler task during the other group's choice phase. The choice groups spent the first portion of their time on the filler task then made final choices (2 × 3) from a base of Six Topics × 12 Items generated and evaluated by the other group. The generation, evaluation, and choice groups spent 60 minutes conducting all three phases of the task. The research assistant was absent from the room during group deliberations. After 20 minutes, he returned to collect decision results to convey to the other group (generation and evaluation), or to convey the other group's inputs to the experimental group (choice), or to provide instructions and forms for evaluation and choice to the generation, evaluation, and choice groups at the appropriate intervals.

Decision Feedback

Guidelines based on the experimental group's input were developed to determine the topics and items in the other group's final choices for feedback to the group in the generation and evaluation conditions. For the generation groups, the following procedure was used. First, from the three topics generated, the highest and lowest risk topics were chosen. Second, one group-generated item was chosen for each of the two topics, along with one nongenerated, slightly higher risk item and one nongenerated slightly lower risk item. Thus, the choice feedback depended, in fact, only partially on the group's generation. For the evaluation groups, the highest and lowest risk topics were chosen, and, for each topic, the group's most favorably evaluated item was included along with two other items that the group had evaluated as slightly unfavorable on the three favorability rating scales.

Decision Risk Measure

Risk scores were directly derived from the items' risk values. For the generation conditions, scores were mean risk scale values for the 18 attitude items generated by the group. For the choice and the generation, evaluation, and choice groups, the means were based on the 6 items chosen during the final choice phase. For the evaluation conditions, the risk means were based only on those items that were evaluated favorably, that is, items that received a total positive rating across all three favorability scales. Unfavorable evaluations were considered analogous to those items in the generation and choice conditions that were, respectively, not generated or chosen.

RESULTS

Manipulation Effectiveness

Both the participation and commitment manipulations were effective. Participation was perceived to have been greater in the partial condition allowing participation in the phase intended by the treatment and in complete participation than in the partial conditions allowing participation in either of the other two phases. In other words, mean perceived participation in the choice phase was significantly greater for the choice (7.86) and generation, evaluation, and choice (8.56) groups than for the generation (3.96) or evaluation (4.07) groups ($F = 36.71$, $df = 3/112$, $p < .001$) and similarly for mean perceived participation in the generation phase ($F = 11.27$, $p < .001$) and for mean participation in the evaluation phase ($F = 12.83$, $p < .001$). Perceived commitment was greater in commitment conditions ($M = 7.75$) than in noncommitment conditions ($M = 2.17$, $F = 210.50$, $df = 1/112$, $p < .001$). No Participation × Commitment interactions significantly affected perceived participation or commitment.

Decision-Making Influence

Results pertinent to the effects of participation and commitment on influence are presented in Table 1. Perceived influence of oneself was greatest with complete participation ($F = 31.65$, $df = 3/112$, $p < .001$, $\eta^2 = .44$).

---

[9] With limited time, the group could not be expected to evaluate all 72 items. Hence, they were given the 18-item subset in topic-item combinations that representatively sampled the range of points along the risk continuum and had nearly equivalent total risk values.

TABLE 1

MEAN INFLUENCE, SATISFACTION, AND DECISION RISK IN PARTICIPATION × COMMITMENT EXPERIMENTAL CONDITIONS

Phase participation	Commitment condition	Dependent measure				
		Self-influence	Others' influence	External influence	Total satisfaction	Decision risk
Generation	C NC	22.00 15.47	40.67 31.00	37.33 53.53	70.27 80.67	3.04 3.07
Evaluation	C NC	17.67 15.13	33.87 28.07	48.47 56.80	77.07 78.67	3.49 3.26
Choice	C NC	33.87 24.67	62.80 46.33	3.33 29.00	79.47 92.13	3.17 2.18
Complete	C NC	37.86 38.40	62.13 61.60	— —	87.07 88.33	2.76 2.27

Note. C = commitment; NC = noncommitment.

Further, when members participated in only one phase, influence was greatest when that phase involved final choices. Orthogonal comparisons among means showed significant differences for complete versus partial participation ($p < .001$) and for the choice phase versus the generation and the evaluation phases ($p < .001$). Influence did not differ between the generation and evaluation conditions. Influence was also significantly greater in the commitment conditions ($F = 6.17$, $df = 1/112$, $p < .05$, $\eta^2 = .03$). The overall commitment effect was attributable to the large difference in choice, which was the only significant difference within participation conditions ($p < .05$).

The influence attributed by the individual to the other two group members and to external groups corroborated this pattern of results. The influence of one's fellow group members was significantly affected by participation ($F = 32.43$, $df = 3/112$, $p < .001$, $\eta^2 = .43$) and by commitment ($F = 9.79$, $df = 1/112$, $p < .01$, $\eta^2 = .04$). As was the case for self-influence, others' influence was greater in complete than in partial participation conditions ($p < .001$), and their influence was also greater in the choice conditions than in the generation or evaluation conditions ($p < .001$). Commitment produced greater influence than noncommitment, especially in choice conditions. The influence of external groups was lowest in complete participation ($F = 54.53$, $df = 3/112$, $p < .001$, $\eta^2 = .55$) and commitment conditions ($F = 14.05$, $df = 1/112$, $p < .001$, $\eta^2 = .05$). Within partial participation conditions, external influence was lowest in the choice phase ($p < .001$).

Influence attributions among group members were also consistently interrelated. First, members agreed on influence attributions. The correlation between Member A's perceived self-influence and the sum of Member B and Member C's attribution of influence to Member A was .73 ($p < .01$). Second, each individual's description of his own influence was positively related to his assessment of the influence exercised by *both* of the other members ($r = .68$, $p < .01$).

Satisfaction

Satisfaction (sum of item ratings) was significantly affected by participation ($F = 4.99$, $df = 3/112$, $p < .01$, $\eta^2 = .11$) and by commitment ($F = 5.92$, $df = 1/112$, $p < .05$, $\eta^2 = .04$). The directions of differences caused by participation were the same as those obtained for the influence variables (Table 1). Satisfaction was greater in complete than partial participation ($p < .02$) and greater in the choice conditions than in the generation or evaluation conditions ($p < .01$). Commitment, on the other hand, affected satisfaction and influence differently.[10] Satisfaction was significantly greater with noncommitment to the decision than under commitment conditions. There was no interaction of this effect with participation treatments. The major commitment difference ($p < .05$) occurred in the choice conditions, where members were much more satisfied if they were not committed to implementing those choices. Satisfaction in the choice–commitment condition was more similar to satisfaction in other partial participation conditions than to complete participation. In contrast, satisfaction in choice–noncommitment was more similar to satisfaction with complete participation, regardless of commitment, than to other forms of partial participation. A similar, but nonsignificant, commitment difference occurred within generation conditions.

Decision Risk

There were no significant effects of either participation or commitment on decision risk. The means in Table 1 indicate that risk was quite low across all conditions (overall $M = 2.91$, $SD = .95$, from a possible range of .5–9.6). Interpretation of these results, which failed to support the hypotheses, requires two additional bits of information. First, the effects appear to depend on perceptions of the importance of the decision task prior to group decision making; that is, pretask attributions of high importance to the task may represent a high degree of personal commitment to the conduct and outcomes of that task prior to experimental manipulations of commitment to implement the decision. This precommitment then tends to enhance the degree to which members strive for high-quality decisions. This situation occurred, as participants' pretask ratings of task importance were uniformly high ($M = 1.97$, $SD = .18$, where a mean of 2.00 indicates that all participants rated the task as important). Moreover, a posttask questionnaire item showed that on a 9-point scale of the degree to which members based their decisions on a quality criterion $M = 7.02$, $SD = 1.70$. Second, if the groups were precommitted to high-quality decisions on the present task and emphasized the quality criterion, the risk of their decisions is necessarily low, since quality and risk were inversely related.

DISCUSSION

Group members' participation in different phases of the decision-making process, and their commitment to the decision, affected perceptions of influence and satisfaction. However, within the constraints of the high-perceived importance of the task, these manipulations did not affect the level of risk in group decision outcomes.

Risk and quality criteria should be reexamined with varying tasks and research designs. For example, with variance in pretask commitment, one could investigate changes in levels of commitment as well as decision risk and quality that result from group decision making. Further work is needed to extend findings on task-specific commitment to the more general hypothesis that participation enhances commitment to organizational goals (Patchen, 1970). Subsequent studies of group decision risk and quality might also compare the effects of individual inputs to decision phases with those of a group process, since the interaction process may differentially affect outcomes of different phases (Vroom, Grant, & Cotton, 1969). Group decision making may also enable disclaiming of individual responsibility or accountability for decisions (Quey, 1971). However, there may be different effects of individuals and groups deciding issues that affect themselves, other members of their task group, or other people who have no role in decision making but for whom the decision makers bear a social responsibility.

The hypothesized effects of participation in decision phases on members' influence and satisfaction were clearly supported. The influence and satisfaction of group members were greatest when members participated throughout the decision process. Moreover, the choice phase was found to yield greater influence than generation or evaluation. These results corroborate earlier evidence that abdication of participation in the choice phase severely limits subordinates' influence (Wood, 1972a). Particularly, if the ideas generated by a group are rejected by others making the choices, participation may generate hostility and alienation (Strauss, 1963). Participation throughout the decision process, and particularly in the final choice of alternatives, appear critical to the sharing of power through group decision making. When members had participated in the complete process, the total group influence pie (Tannenbaum, 1968) was largest. Moreover, members' assessments of their own influence were positively related to their assessments of the influence of the other members of their group. Thus, the enlarged pie was shared among all group members, rather than being attributable to the greater influence of one person at the expense of others.

---

[10] Perceived self-influence and satisfaction were generally independent. The only satisfaction item correlating significantly with influence was the measure of members' satisfaction with final attitude items chosen to be included in the survey ($r = .44$, $p < .01$). See Wood (1972b) for possible motivational moderators of influence–satisfaction relationships.

It must be noted that the present groups were leaderless. Previous research has shown that when groups begin the decision process with a legitimately defined leader, complete participation enhances the influence pie by increasing members' influence without detracting from (but not necessarily increasing) the leader's influence (Wood, 1972a). Further research on intact organizational groups over time, where leaders and subordinates are jointly affected by decision feedback, is needed to clarify the mechanisms of influence sharing and its effects on organizational behavior.

Participative management has been generally shown to relate to the sharing of influence across organizational levels and functions and to favorable organizational outcomes (Likert, 1967). The present research indicates that these relationships (and, more importantly, the effects of participation) depend on the specific manner in which the participative management approach is applied to the process of making organizational decisions.

Group members' commitment to the decision task also affected influence and satisfaction, although this treatment accounted for much less variance in the dependent measures than did participation. Influence effects were in the predicted direction where commitment generated greater influence. However, noncommitment led to greater satisfaction, especially when members participated only in choosing alternatives generated and evaluated by others. A post hoc interpretation of this finding may be based on the assumption of less pressure toward compliance in the evaluative or analytic phase of decision making. In contrast to generation or choice, the open group discussion and evaluation of alternatives, which the members have not produced and do not have to adopt, may generate as much satisfaction when members are committed to implementing the decision as when they are not (Doyle, 1971). Future research should study additional situational parameters that may affect, or interact with commitment to affect, outcomes of group decision making.

REFERENCES

BELOVICZ, M. W., & FINCH, F. E. A critical analysis of the "risky shift" phenomenon. *Organizational Behavior and Human Performance*, 1971, **6**, 150-168.

CLARK, R. D., III, & WILLEMS, E. P. Risk preferences as related to judged consequences of failure. *Psychological Reports*, 1969, **25**, 827-830.

COOPER, M. R. *Satisfaction, perceived influence, and decision-riskiness: The effects of participation and commitment in decision-making groups.* (Doctoral dissertation, Ohio State University) Ann Arbor, Mich.: University Microfilms, 1972, No. 72-26,994.

DION, K. L., MILLER, N., & MAGNAN, M. A. Cohesiveness and social responsibility as determinants of group risk taking. *Journal of Personality and Social Psychology*, 1971, **20**, 400-406.

DOYLE, W. J. Effects of achieved status of leader on productivity of groups. *Administrative Science Quarterly*, 1971, **16**, 40-50.

JOHNSON, D. L., & ANDREWS, I. R. Risky-shift phenomenon tested with consumer products as stimuli. *Journal of Personality and Social Psychology*, 1971, **20**, 382-385.

LIKERT, R. *The human organization: Its management and value*. New York: McGraw-Hill, 1967.

MANN, L. Effects of a commitment warning on children's decision behavior. *Journal of Personality and Social Psychology*, 1971, **17**, 74-80.

PATCHEN, M. *Participation, achievement, and involvement on the job*. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

PRUITT, D. G. Choice shifts in group discussion: An introductory review. *Journal of Personality and Social Psychology*, 1971, **20**, 339–358.

QUEY, R. L. Functions and dynamics of work groups. *American Psychologist*, 1971, **26**, 1077–1082.

STRAUSS, G. Some notes on power-equalization. In H. J. Leavitt (Ed.), *The social science of organization*. Englewood Cliffs, N.J.: Prentice-Hall, 1963.

TANNENBAUM, A. S. *Control in organizations*. New York: McGraw-Hill, 1968.

VROOM, V. H., GRANT, L. D., & COTTON, T. S. The consequences of social interaction in group problem solving. *Organizational Behavior and Human Performance*, 1969, **4**, 77–96.

WOOD, M. T. Effects of decision processes and task situations on influence perceptions. *Organizational Behavior and Human Performance*, 1972, **7**, 417–427. (a)

WOOD, M. T. Participation, influence, and satisfaction in group decision making. *Journal of Vocational Behavior*, 1972, **2**, 389–399. (b)

WOOD, M. T. Power relationships and group decision making in organizations. *Psychological Bulletin*, 1973, **79**, 280–293.

(Received January 4, 1973)

Manuscripts Accepted for Publication in the
Journal of Applied Psychology

Development of Scales to Measure Attitudes toward Marijuana and Marijuana Users. Victoria Colaiuta and George Breed (Department of Psychology, University of South Dakota, Vermillion, South Dakota 57069).

A Dynamic Model of the Relationships among Advertising, Consumer Awareness, Attitudes, and Behavior. David A. Aaker (School of Business Administration, University of California, Berkeley, California 94720) and George S. Day.

Equal-Unpleasantness Matching of Malodorous Substances in the Community. Leif T. Svensson (Psychological Laboratories, University of Stockholm, Box 6706, S-113 85, Stockholm, Sweden) and Thomas Lindvall.

Magnitude Estimations of Expressions of Frequency and Amount. Bernard M. Bass (Management Research Center, Graduate School of Management, University of Rochester, New York 14627), Wayne F. Cascio, and Edward J. O'Connor.

A Note on the Predictive Validity of Initiating Structure and Consideration in ROTC Training. M. M. Petty (Manpower Industrial Relations and Organizational Behavior Program, University of Alabama, P.O. Box 2737, Alabama 35486) and Norman M. Pryor.

Contribution of Personalization to Mail Questionnaires Response as an Element of a Previously Tested Method. Don A. Dillman (Department of Rural Sociology, Washington State University, Pullman, Washington 99163) and James H. Frey.

Some Moderator Influences on the Relationships between Consideration, Initiating Structure, and Organizational Criteria. Irwin J. Badin (Psychology Department, Montclair State College, Upper Montclair, New Jersey 07043).

Visual Dependency in the Erect and Supine Positions. Jacob H. Lichtenstein (Flight Dynamics and Control Division, National Aeronautics and Space Administration, Langley Research Center, Hampton, Virginia 23365) and Rayford T. Saucer.

Effects of Different Levels of Integration on Advertising Preference and Intention to Purchase. George J. Szybillo (Graduate School of Business Administration, New York University, 100 Trinity Place, New York, New York 10006) and Jacob Jacoby.

Evaluation of a Deceleration Warning Light for Reducing Rear-End Automobile Collisions. John Voevodsky (177 Goya Road, Portola Valley, California 94025).

Probability and Utility Assumptions Underlying Use of the Strong Vocational Interest Blank. Frank L. Schmidt (Department of Psychology, Olds Hall, Michigan State University, East Lansing, Michigan 48823).

Relationship between Managerial Values and Managerial Success in the United States, Japan, India, and Australia. George W. England (Industrial Relations Center, 537 Business Administration Building, University of Minnesota, Minneapolis, Minnesota 55455) and Raymond Lee.

Effects of Applicant's Sex and Difficulty of Evaluations of Candidates for Managerial Positions. Benson Rosen (Graduate School of Business Administration, University of North Carolina, Chapel Hill, North Carolina 27514) and Thomas H. Jerdee.

(*Continued on page 144*)

Journal of Applied Psychology
1974, Vol. 59, No. 2, 135–138

BRAINSTORMING PROCEDURE, GROUP SIZE, AND SEX AS DETERMINANTS OF THE PROBLEM-SOLVING EFFECTIVENESS OF GROUPS AND INDIVIDUALS[1]

THOMAS J. BOUCHARD, JR.,[2] JEAN BARSALOUX, AND GAIL DRAUDEN

University of Minnesota

A $2 \times 2 \times 2 \times 2$ analysis of variance was used to assess: (*a*) a modified brainstorming procedure in which a physical environmental influence was concordant with the task versus standard brainstorming, (*b*) sex of group (males versus females), (*c*) size of group (four persons versus seven persons), and (*d*) type of group (real versus nominal). The effect of the brainstorming treatments on males and females as individuals was also assessed. The environmental concordance treatment reduced the effectiveness of the nominal groups, but had no effect on real groups or individuals. Sex interacted with treatment, group size and type of group. Seven-person, nominal groups outperformed four-person, nominal groups, but there was no size effect for real groups. Nominal groups outperformed real groups.

In a previous study Bouchard (1972a) contrasted two group problem-solving procedures—standard brainstorming and a modification of brainstorming which forced greater involvement of the participants. The modified procedure made use of a tactic called "personal analogy" (cf. Gordon, 1961) which required the participants to play the role of an object or concept central to the problem being worked on. This was called the synectics condition. The synectics treatment was shown to be more effective than was standard brainstorming on a variety of divergent thinking tasks. The procedures used in that study, however, confounded active involvement with the task and object role playing or personal analogy. Further, for one of the tasks the experimenters also varied an important environmental parameter in the synectics condition. The task was the "blindness problem." The problem required the subjects to come up with a list of consequences that would occur if suddenly everyone went blind. During the session, each of four subjects was, in turn, required to wear a blindfold for 3 minutes. At the end of 12 minutes the lights were turned out, and the subjects worked an additional 8 minutes in the dark. Turning the lights off during work on the blindness problem constitutes what we call a Problem × Environment concordance.

This experiment is an attempt to further investigate the factors discussed above and to assess treatment generality across sex, type of group (real, nominal, individuals), and size of group. This study is restricted to testing the hypothesis that Problem × Environment concordance alone is a sufficiently powerful treatment to generate an increase in performance over normal problem-solving procedures. The treatment does not require either active involvement or role playing. The hypothesis has a plausible basis in that there are numerous anecdotal reports that creative discoveries have occurred when the problem solver put or found himself in the context of the problem. The classic example is that of Archimedes who is claimed to have discovered how to measure the volume of a gold crown (by measuring its displacement of water) while sitting in a bathtub. Koestler (1954) and Krippner and Hughes (1970) cite additional examples.

The experiment is a four-way analysis of variance designed to assess the treatment effects on two different-sized groups (four-

---

[1] This investigation was supported entirely by Public Health Service Grant 1 R01 HD 5600-01 to the first author. The report was written while the first author was at the Oregon Research Institute supported by General Research Support Grant RR-05612 from the National Institutes of Health and Grant MH-12972 from the National Institute of Mental Health, United States Public Health Service.

[2] Requests for reprints should be sent to Thomas J. Bouchard, Jr., Elliott Hall, Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455.

TABLE 1

MEANS AND STANDARD DEVIATIONS FOR SIMPLE EFFECTS AND VARIOUS MAIN EFFECTS FOR TOTAL NUMBER OF IDEAS

Measure	Group size four								Group size seven							
	Nominal				Real				Nominal				Real			
	Light		Dark		Light		Dark		Light		Dark		Light		Dark	
	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
Simple effects																
M *SD*	283.5 6.5	337.0 15.0	249.0 6.0	217.0 8.0	140.0 4.0	109.0 22.0	144.0 10.0	113.0 22.0	415.0 8.0	454.0 7.0	511.5 41.5	264.0 8.0	153.0 16.0	149.5 15.5	149.5 28.5	144.0 5.0
Light vs. dark																
M *SD*	310.3 29.1		233.0 17.5		124.5 22.1		128.5 23.0		434.5 20.9		387.8 127.3		151.3 15.8		146.8 20.6	
Nominal vs. real																
M *SD*	271.6 45.5				126.5 22.7				411.1 94.17				149.0 18.54			

and seven-person groups), groups of both sexes, and real versus nominal groups. Treatment effects on individual problem solvers were also assessed.

METHOD

Subjects

The subjects were male and female students from an introductory psychology course. They were recruited to the experiment by offering them points for experimental participation (required of all introductory psychology students).

Design

The experimental design, a $2 \times 2 \times 2 \times 2$ analysis of variance is shown in Table 1, where the data for single main effects are presented. There were two groups per cell.

Procedure

One experimental unit in each cell was run by each of the second and third authors (females). All of the instructions were presented by tape recorder. Each subject also had a written copy of the rules of brainstorming to examine while he listened to the instructions. A brief summary is given below.[3]

Group brainstorming instructions. The first part of the instructions were identical to those used by Bouchard and Hare (1970) for groups. These were followed by the forced-participation procedure described below. The brainstorming procedure proscribed criticism and emphasized free-wheeling, quantity, and the combination and improvement of ideas. The forced-participation procedure required ideas to be contributed by participants in sequence. Each subject had to say "pass" if he had nothing to say. Subjects were also encouraged not to get bogged down in long discussions about trivial points.

When the instructions were over the experimenter returned and answered any questions. Subjects were then told "Here is the problem you will be working on." They were then handed the problem. If the group was under the lights-on condition, the problem read as follows: "What would be the consequences if suddenly everybody went blind? You will have thirty-five (35) minutes to work on the problem. The experimenter will tell you when to stop."

If the group was under the lights-off condition, the problem read as follows:

> What would be the consequences if suddenly everyone went blind? You will begin working on this problem in dim light. During the course of the task the experimenter will decrease the illumination until it is pitch black. If for some reason you must have the light back on, just say so and the experimenter will put it on. This will disrupt the experiment so please don't ask unless it is necessary. You will have thirty-five (35) minutes to work on the problem. The experimenter will tell you when to stop.

---

[3] Complete instructions are available directly from the author.

The subjects were given time to read the problem and then told: "You will have thirty-five minutes to work on it. I will tell you when the time is up. Do you have any questions about the problem? OK, you can begin."

The darkness groups worked for the first five minutes with a 60-watt desk lamp shining against the wall. During this time the room was dim, but it was possible to see all group members and make out facial expressions. During the second five minutes the lamp was extremely dim and served simply to give stability to the situation. It was turned down slowly with a rheostat. During this period the subjects could not see each other even after dark adaptation. After the second five minutes, the lamp was turned off completely and the room was absolutely pitch black.

Individual brainstorming conditions. The first part of the instructions were identical to those used by Bouchard and Hare (1970) for individuals. The remainder of the instructions were essentially identical to those used in the groups.

A single microphone sat on the table in front of individual subjects, while four microphones were used for the various group conditions.

Scoring

The dependent variable analyzed was the number of different ideas generated. Previous work (Bouchard, 1969, 1972b) has shown that the total number of ideas is the most sensitive indicator of treatment effects and correlates highly with quality scores. For the real groups, all overlapping ideas within a group were deleted from the transcribed tapes. The remaining number of ideas constitute a fluency score for the group. For the nominal groups, a fluency score was created by counting the nonoverlapping ideas of four individuals who worked alone in the appropriate condition. Interjudge reliability of the deletion procedure is generally very high and only one judge was used (cf. Bouchard & Hare, 1970).

RESULTS

The means and standard deviations for simple effects are given in Table 1. The analysis of variance is given in Table 2. All of the main effects were significant, but so were most of the interactions, and this tempered any interpretation of main effects.

Treatment (Light versus Dark) and Type of Group (Nominal versus Real)

The significant Type × Treatment and Treatment × Sex interactions can be easily interpreted when the significant Type × Treatment × Sex interaction is plotted.[4] The triple interaction (as well as the four-way interaction) occurred because the real groups were not responsive to any experimental manipulations, while the nominal groups were. The nominal groups were, however, far superior to the real group (341.4 versus 137.8). Focusing on the nominal group data only, the Sex × Treatment interaction was highly significant ($F = 61.06$, $df = 1/8$, $p < .01$). Women performed better than men under the lights-on condition, while men performed better than women in the lights-off condition. The slight improvement from the lights-on condition to the lights-off condition for the male nominal groups occurred only for the group with seven subjects and did not appear to be a meaningful treatment effect. The lights-off treatment depressed the performance of the nominal groups relative to the lights-on treatment (310.4 versus 372.4; $F = 27.14$, $df = 1/8$, $p < .01$). Over all conditions, the lights-on treatment was superior to the lights-off treatment (255.13 versus 224.0; $F = 13.24$, $df = 1/16$, $p < .01$).

TABLE 2

ANALYSIS OF VARIANCE OF MEAN TOTAL NUMBER OF IDEAS PRODUCED UNDER EACH CONDITION

Source	*df*	*MS*	*F*
Size (A)	1	52,488.0	89.65**
Type (B)	1	331,705.1	566.53**
Treatment (C)	1	7,750.1	13.24**
Sex (D)	1	8,320.5	14.21**
A × B	1	27,378.0	46.76**
A × C	1	242.0	
A × D	1	3,916.1	6.69*
B × C	1	7,626.1	13.02**
B × D	1	1,682.0	2.87
C × D	1	17,484.5	29.86**
A × B × C	1	760.5	1.30
A × B × D	1	10,011.1	17.10**
A × C × D	1	5,151.1	8.80**
B × C × D	1	17,112.5	29.29**
A × B × C × D	1	4,950.1	8.45*
Within replicates	16	585.5	
Total			

* $p < .05$.
** $p < .01$.

Group Size

The significant Size × Type and Size × Sex interactions can be easily interpreted when the Size × Type × Sex interaction is plotted (see Footnote 4). The nominal groups were superior to the real groups, and nominal group performance increased with group size, especially for males, while real group performance were not. The Size × Sex interaction were highly significant when the nominal groups were analyzed alone ($F = 23.34$, $df = 1/8$, $p < .01$).

---

[4] The reader can easily plot all interaction from the data presented in Table 1.

Individual Performance

In order to construct the nominal groups, 44 males and 44 females were run individually, half of each sex with the lights on and the other half with the lights off. There were no treatment effects, sex differences, or interaction effects. The male and female means under the light conditions were 97.00 and 100.05. Under the dark conditions, they were 98.45 and 112.59.

Discussion

The dependent variable was shown to be responsive to a variety of factors (sex, size of group, type of group); thus, the significant decrease in the performance of the nominal groups under the lights-off treatment and the failure to detect any treatment effect under the individual conditions suggest that the Problem × Environment concordance hypothesis as formulated in this study is false.[5] Creating a passive congruence between the task and the environment does not increase productivity.

The real group versus nominal group comparison allowed a suggestion by Bouchard (1972b) to be tested, namely, that the forced participation (used in this study) might lift the performance of real groups to the level of nominal groups. That suggestion was clearly disconfirmed. The forced-participation effect found in the Bouchard (1972b) study must have been due to the special nature of the groups used. They were all low scorers on a personality factor called interpersonal effectiveness.

The Size × Type effects are a striking replication, using a new problem, new procedures, and a longer working time, of results previously obtained by Bouchard and Hare (1970). Increasing real group size adds very little to performance, while increasing nominal group size increases performance dramatically.

REFERENCES

Bouchard, T. J., Jr. Personality, problem-solving procedure, and performance in small groups. *Journal of Applied Psychology,* 1969, **53**, 1–29.

Bouchard, T. J., Jr. A comparison of two group brainstorming procedures. *Journal of Applied Psychology,* 1972, **56**, 418–421. (a)

Bouchard, T. J., Jr. Training, motivation, and personality as determinants of the effectiveness of brainstorming groups and individuals. *Journal of Applied Psychology,* 1972, **56**, 324–331. (b)

Bouchard, T. J., Jr., & Hare, M. Size, performance, and potential in brainstorming groups. *Journal of Applied Psychology,* 1970, **54**, 51–55.

Gordon, W. J. *Synectics: The development of creative capacity.* New York: Harper & Row, 1961.

Koestler, A. *The act of creation.* New York: Dell, 1964.

Krippner, S., & Hughes, W. ZZZ genius at work ZZZ. *Psychology Today,* 1970, **4**, 40–43.

(Received February 12, 1973)

---

[5] We have also failed to confirm the Problem × Environment concordance hypothesis in two previous experiments. Technical reports are available from the first author.

Journal of Applied Psychology
1974, Vol. 59, No. 2, 139-144

A CAUSAL-CORRELATIONAL ANALYSIS OF THE JOB SATISFACTION AND PERFORMANCE RELATIONSHIP

JOHN P. WANOUS[1]

Graduate School of Business Administration, New York University

Job satisfaction and performance data were gathered from about 80 newly hired female telephone operators after one and three months work experience. This was not an experimental study, but tentative causal inferences were drawn from these longitudinal data by using a combination of cross-lagged and dynamic correlations. The overall relationship between satisfaction and performance was slightly positive, but the direction of causality was unclear. When job satisfaction was split into extrinsic and intrinsic components, the data suggest that performance causes intrinsic satisfaction and that extrinsic satisfaction causes performance. Both the type of job and the fact that these were new hires were offered as alternative explanations for these results.

Prior to the Hawthorne research (Roethlisberger & Dickson, 1939), job performance probably was the major dependent variable studied by industrial researchers. During the human relations era following Hawthorne, the view of work environments expanded to include employee attitudes, especially job satisfaction. It was during this period that a causal relationship between satisfaction and performance was assumed, that is, high satisfaction leading to high performance.

Literature reviews in the middle 1950's (Brayfield & Crockett, 1955; Herzberg, Mausner, Peterson, & Capwell, 1957) cast doubt on both the form, strength and causal direction of the satisfaction and performance relationship. Although Brayfield and Crockett were more pessimistic than Herzberg et al., both reviews illuminated the conflicting results of empirical research. The net effect of these reviews, especially Brayfield and Crockett's, resulted in what Schwab and Cummings (1970) have called "the development of uncertainty" or "satisfaction—?—performance." Uncertainty was not the last phase in this research, however. A more recent view is that performance leads to satisfaction, whereby performance results from the joint effects of effort, ability, and role perceptions (Porter & Lawler, 1968).

Recent Literature on the Satisfaction and Performance Relationship

In the early 1970's alone both theoretical statements (Locke, 1970; Sutermeister, 1971) and empirical research (Cherrington, Reitz, & Scott, 1971; Siegel & Bowen, 1971; Slocum, 1970, 1971) have been published, attesting to continued interest in this topic.

Locke (1970) has suggested that satisfaction is primarily a result of performance and only indirectly a cause of performance. The strength of this relationship is said to depend on ". . . the degree to which performance entails or leads to the attainment of the individual's important job values (without neglecting his other important values) [1970, p. 485]." Locke found that emotions (such as satisfaction) do not usually determine behavior, because job behavior results from an individual's own task goals.

Sutermeister (1971) has stated a cyclical model of the satisfaction and performance relationship. Life style ("climber, conserver, or slider") is viewed as an important determinant of level of aspiration, which, in turn, affects individual effort in the work situation. The cyclical model states that satisfaction and performance are causes of each other, however the closer (and presumably stronger) linkage runs from performance to satisfaction. Thus, the theoretical positions of Porter and Lawler, Sutermeister, and Locke are all in agreement, that is, performance is a much stronger cause of satisfaction than the reverse. This view seems to prevail today.

---

[1] The author wishes to thank J. Richard Hackman, Abraham K. Korman, Edward E. Lawler III, and Benjamin Schneider for their comments.

Requests for reprints should be sent to John P. Wanous, Graduate School of Business Administration, New York University, 100 Trinity Place, New York 10006.

Several recent empirical studies have also been published. In two articles, using the same sample of about 200 lower and middle level supervisors, Slocum (1970, 1971) analyzed this relationship using concurrent (not predictive) correlations. Slocum's first study focused on satisfaction/performance correlations for each different need level. This was a test of the view that performance and satisfaction were more closely related for higher than lower need levels. A significantly higher correlation was found for self-actualization needs than for either security or esteem needs, similar to the findings of an earlier study by Lawler and Porter (1967). In the second article Slocum's study focused on satisfaction/performance correlations for different managerial levels. Performance and satisfaction were found to be more highly correlated for upper as opposed to lower level managers.

Siegel and Bowen (1971) gathered longitudinal data, permitting an analysis of both concurrent and predictive relationships. Their sample was 86 M.B.A. students from a classroom situation where performance and satisfaction measures were based on course work. They found strongest support for performance as the cause of satisfaction.

Cherrington, Reitz, and Scott (1971) conducted an experiment in which student subjects were either randomly, appropriately, or inappropriately rewarded for performance. Depending on the situation an individual was in, the satisfaction/performance correlations were either zero (random rewards), positive (contingent rewards appropriate for performance), or negative (contingent rewards inappropriate for performance).

The Present Research

Although many studies have focused on this topic, the present effort appears to be different. First, multiple measures of both satisfaction and performance were gathered and combined in composite form. Two widely accepted job satisfaction scales were used. Second, both cross-lagged and dynamic correlations were computed. Third, job satisfaction was split into extrinsic and intrinsic composites and each was separately correlated with performance.

METHOD

This research was conducted in an eastern telephone company, using a sample of about 80 newly hired female operators who volunteered. Questionnaires designed to measure a number of attitudes toward the job were administered after one and three months working experience; only the job satisfaction attitudes are reported in this article.

Three measures of job satisfaction were gathered: (*a*) the Job Descriptive Index (JDI), (*b*) the Minnesota Satisfaction Questionnaire, Short Form (MSQ), and (*c*) two items designed to measure general satisfaction. One of these direct global items asked for satisfaction to be rated directly on a 5-point Likert-type scale. The other item asked subjects to rate on a 5-point scale how the characteristic "being very satisfied" was for them at work. The correlation between these two items was .76 at one month and .80 at three months. Although Wanous and Lawler (1972) have shown that job satisfaction has been measured in nine different ways, only well known, direct measures were used in this study.

The MSQ is used in two ways, to measure overall satisfaction and to measure intrinsic versus extrinsic satisfaction. Previous researchers (Weiss, Dawis, England, & Lofquist, 1967) intended that the sum of all 20 items reflect global job satisfaction and used factor analysis to derive extrinsic and intrinsic factors. The JDI is also used to measure overall job satisfaction, although this was not its intent as originally designed (Smith, Kendall, & Hulin, 1969). However, the sum of all five scales of the JDI has been used at least three times prior to this research (Ewen, 1967; Hulin, 1968; Mikes & Hulin, 1968), and in the present case correlated .71 at one month and .72 at three months with the MSQ total. Thus, it is felt that the JDI total does have meaning and can be used as a global measure of satisfaction. To measure extrinsic versus intrinsic satsisfaction, the JDI was split into its Work and Promotions scales versus Pay, Co-workers, and Supervision, consistent with Herzberg (1959).

Corresponding to the multiple measurement of job satisfaction, several measures of job performance were gathered at one and three month intervals: supervisory ratings of employee quality and quantity of performance, as well as company indexes of quality and quantity of performance.

Because multiple measures of both satisfaction and performance were obtained, it was decided to combine them into composites for simplified analysis. To form a composite of overall job satisfaction, the

JDI, MSQ, and global items were each first treated as separate scales. Only the totals of each scale were used. Because of widely differing variances for each satisfaction scale, each was standardized before being combined to form the composite of overall job satisfaction. A similar procedure was followed for combining supervisory ratings with company indexes to form a composite of overall job performance.

Following analysis of the overall satisfaction and performance relationship, satisfaction was split into its extrinsic and intrinsic components. The performance composite was not changed but the satisfaction composite was changed as follows: First, the direct global items were dropped. Second, the JDI and MSQ were split into extrinsic and intrinsic components, as described previously in the study. Each component was then standardized before combining the JDI and MSQ into intrinsic and extrinsic components.

Data analysis involved two types of correlational procedures: (*a*) cross-lagged (e.g., see Lawler, 1968) and (*b*) dynamic (Vroom, 1966). Although causality cannot be proven by the combination of these two procedures, possible causal relationships may be suggested. The cross-lagged method is used when two variables are measured at two points in time, so that both concurrent and predictive relationships between all pairs of variables are represented by correlations.

If performance is the cause of satisfaction, then one expects to find a pattern of correlations something like $r_6 > r_1 = r_3 > r_5$. One problem with this approach is that few theories of organizational behavior specify time lags for causal effects. In the present case a lag of two months was used, but this may or may not be appropriate for possible causal effects to be detected on the two variables studied in this research.

Another weakness of cross-lagged correlational analysis is that it may show significant results when some third variable has influenced the observed relationship between performance and satisfaction. In an attempt to deal with the weakness, Vroom (1966) proposed the use of dynamic correlations. A dynamic correlation relates *changes* in one variable to changes in the other, for example, performance and satisfaction. If the pattern of dynamic correlations is similar to the pattern of cross-lagged correlations, it is unlikely that a third variable caused the observed relationship. For a third variable to cause a spurious correlation between two difference scores, performance change and satisfaction change, it must change in different amounts or directions in the members of the sample, and these changes must be highly correlated with both performance changes and satisfaction changes. In the present study, dynamic correlations were computed by using the changes in job satisfaction and performance composite scores between one and three months on the job.

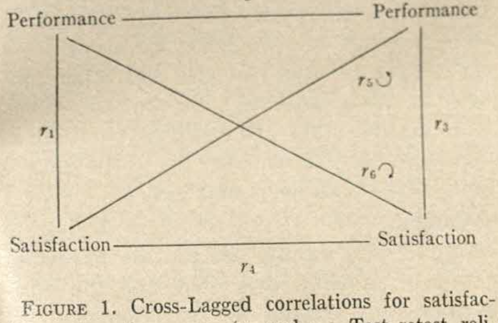


FIGURE 1. Cross-Lagged correlations for satisfaction and performance. (r_2 and r_4: Test–retest reliabilities; r_1 and r_3: Concurrent relationships; r_5 and r_6: Predictive relationships.)

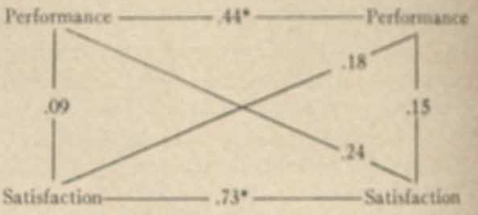


FIGURE 2. Overall job satisfaction and performance relationship. (*$p < .01$).

RESULTS

The results of various cross-lagged correlations are presented in Figures 2, 3, and 4; dynamic correlations are shown in Table 1. The overall job satisfaction and performance relationship in Figure 2 shows all positive correlations. However, only the test–retest correlations for job satisfaction and performance were significant. None of the concurrent or predictive satisfaction and performance correlations were significant.

When job satisfaction was split into extrinsic and intrinsic composites,[2] a slightly different pattern of correlations emerged.

TABLE 1

DYNAMIC CORRELATIONS

Performance	Satisfaction		
	Overall	Intrinsic	Extrinsic
	.24	.30*	.21

* $p < 05$.

---

[2] The correlations between intrinsic satisfaction and extrinsic satisfaction were .18 (*ns*) at one month and .23 (*ns*) at three months.

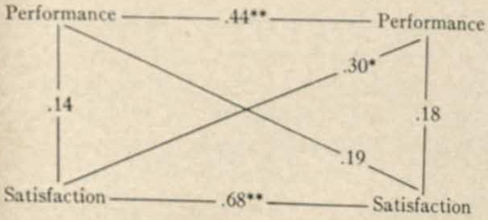


FIGURE 3. Extrinsic satisfaction and job performance. (* $p < .05$; ** $p < .01$).

Figure 3 shows the relationship between extrinsic satisfaction and job performance. Again, all correlations were positive, and satisfaction and performance were most strongly related to themselves. In contrast to Figure 2, however, a significant correlation appeared between extrinsic satisfaction at one month and job performance at three months. Although this particular relationship was statistically significant, it was not significantly different from any of the other three correlations relating satisfaction and performance.

When job satisfaction was intrinsic, as shown in Figure 4, the correlations were all positive. Again, satisfaction and performance correlated most highly with themselves. The results in this study show that performance at one month was significantly related to subsequent intrinsic job satisfaction. This result was just the reverse of the findings when extrinsic satisfaction was related to performance. Not only was the $r = .37$ significant, but it is significantly different from both concurrent correlations.[3] The difference between the $r = .37$ and the two cross-lagged correlations was only close to significance ($p < .09$).

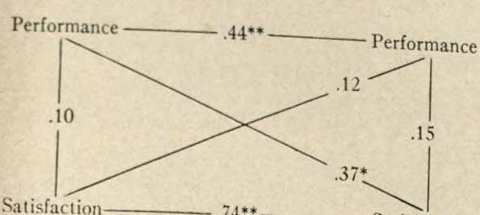


FIGURE 4. Intrinsic satisfaction and job performance. (* $p < .05$; ** $p < .01$).

Table 1 shows the dynamic correlations corresponding to each of the Figures 2, 3, and 4. Although all three were positive, only the intrinsic satisfaction and job performance correlation was significant. This result was parallel to the findings in Figure 4. In fact, finding a positive, nonsignificant correlation for overall satisfaction and performance also paralleled the results in Figure 2. The dynamic correlation between extrinsic satisfaction and performance was not significant and did not lend support to the results in Figure 3.

DISCUSSION

Theory and research concerning job satisfaction and performance date back a number of years in studies of organizational behavior. Although some have called for a moratorium on research of this relationship (Schwab & Cummings, 1970), it was felt that some improvement might be made over previous research in this area. Using both theoretical and methodological contributions from recent years, a more comprehensive strategy for analysis was designed which employed composites of satisfaction and performance based on multiple measures, the distinction between intrinsic and extrinsic satisfaction, and the combination of cross-lagged and dynamic correlations.

The results are somewhat similar to previous research, as well as suggestive of new directions for theory and research. Any conclusions drawn from this research must, however, remain somewhat tentative. This is because of limitations on subject and job samples used in this study. These data were based on females only and were limited to those working as telephone operators. Generalizations to other samples and other jobs must be made cautiously. On the other hand, it is possible that relationships reported in this study were attenuated by restrictions in range of both subjects and jobs.

Caution must be exercised in trying to interpret these results because the particular nature of subjects, their jobs, and the organization may account for the differences between extrinsic and intrinsic satisfaction. It must be kept in mind that this sample was all newly hired, had never been telephone operators previously, and was entirely female.[4] It is possible to explain the relationship between initial extrinsic satisfaction and subsequent performance in terms of a desire to "look good" to supervisors. This seems to be reasonable as these people were all new to the organization. Perhaps the human relations' view ("a satisfied worker is a good worker.") is a correct description of new employee motivation.

---

[3] When testing for a significant difference between .37 and the two concurrent *r*s, .10 and .15, a *t* test which accounts for the common array was used. The results showed that .37 is significantly different from .10 ($p < .01$) and from .15 ($p < .05$).

It is also possible to interpret the initial performance and subsequent intrinsic satisfaction relationship as a result of this particular situation. The telephone operators' job is not one with potential for satisfying "higher order" needs or being intrinsically satisfying (Hackman & Lawler, 1971; Wanous, 1973). However, it is often experienced as initially challenging. The monotony, lack of variety and low autonomy are not typically experienced by newly hired operators. Thus, it is possible that the significant correlation between initial performance and subsequent intrinsic satisfaction found in this study might not appear in a sample of more experienced employees.

The results and discussion indicate that there probably is no single "correct" relationship between satisfaction and performance. Sometimes there appears to be no relationship at all, which is consistent with the view that situations can be created by reward systems that will support any causal model. However, the view taken in this research is that there are additional conditions besides reward systems when certain performance and satisfaction relationships are significant, for example, type of satisfaction, work experience of individuals, and type of job situation they are in. This research should be considered as a first step in the identification of such conditions and in the empirical mapping of possible causal relationships.

Future research and thinking might consider further the occasions when performance seems to be the cause of satisfaction and vice versa, as suggested by the present study. The data in this research are in general agreement with a reciprocal causation view of job satisfaction and performance, as recently stated several times. To elaborate such a view of satisfaction and performance, more research is needed which specifies in detail the effects of different job settings, subject samples, and measures of performance and satisfaction on the causal and feedback relationships between satisfaction and performance.

REFERENCES

Brayfield, A. H., & Crockett, W. H. Employee attitudes and employee performance. *Psychological Bulletin*, 1955, **52**, 396–424.

Cherrington, D. J., Reitz, H. J., & Scott, W. E. Effects of contingent and noncontingent reward on the relationship between satisfaction and task performance. *Journal of Applied Psychology*, 1971, **55**, 531–536.

Ewen, R. B. Weighting components of job satisfaction. *Journal of Applied Psychology*, 1967, **51**, 68–73.

Hackman, J. R., & Lawler, E. E. Employee reactions to job characteristics. *Journal of Applied Psychology*, 1971, **55**, 259–286.

Herzberg, F., Mausner, B., Peterson, R. O., & Capwell, D. F. *Job attitudes: A review of research and opinion*. Pa.: Psychological Service of Pittsburgh, 1957.

Herzberg, F., Mausner, B., & Snyderman, B. *The motivation to work*. New York: Wiley, 1959.

Hulin, C. L. Effects of changes in job satisfaction levels on employee turnover. *Journal of Applied Psychology*, 1968, **52**, 122–126.

Lawler, E. E. A correlational–causal analysis of the relationship between expectancy attitudes and job performance. *Journal of Applied Psychology*, 1968, **52**, 462–468.

Lawler, E. E., & Porter, L. W. The effect of performance on job satisfaction. *Industrial Relations*, 1967, **7**, 20–28.

Locke, E. A. Job satisfaction and job performance: A theoretical analysis. *Organizational Behavior and Human Performance*, 1970, **5**, 484–500.

Mikes, P. S., & Hulin, C. L. Use of importance as a weighting component of job satisfaction. *Journal of Applied Psychology*, 1968, **52**, 394–398.

Porter, L. W., & Lawler, E. E. *Managerial attitudes and performance*. Homewood, Ill.: Richard D. Irwin, 1968.

Roethlisberger, F. J., & Dickson, W. J. *Management and the worker*. Cambridge, Mass.: Harvard University Press, 1939.

Schwab, D. P., & Cummings, L. L. Employee performance and satisfaction with work roles: A review and interpretation of theory. *Industrial Relations*, 1970, **9**, 408–430.

Siegel, J. P., & Bowen, D. Satisfaction and per-

---

[4] The sample was, however, reasonably "typical" in its distribution of age, race, and educational background.

formance: Causal relationships and moderating effects. *Journal of Vocational Behavior*, 1971, **1**, 263–269.

SLOCUM, J. W. Performance and satisfaction: An analysis. *Industrial Relations*, 1970, **9**, 431–436.

SLOCUM, J. W. Motivation in managerial levels: Relationship of need satisfaction to job performance. *Journal of Applied Psychology*, 1971, **55**, 312–316.

SMITH, P. C., KENDALL, L. M., & HULIN, C. L. *The measurement of satisfaction in work and retirement*. Chicago, Ill.: Rand McNally, 1969.

SUTERMEISTER, R. A. Employee performance and employee need satisfaction—Which comes first? *California Management Review*, 1971, **13**, 43–47.

VROOM, V. A. A comparison of static and dynamic correlational methods in the study of organizations. *Organizational Behavior and Human Performance*, 1966, **1**, 55–70.

WANOUS, J. P. Effects of a realistic job preview on job acceptance, job attitudes, and job survival. *Journal of Applied Psychology*, 1973, **58**, 327–332.

WANOUS, J. P., & LAWLER, E. E. Measurement and meaning of job satisfaction. *Journal of Applied Psychology*, 1972, **56**, 95–105.

WEISS, D. J., DAWIS, R. V., ENGLAND, G. W., & LOFQUIST, L. H. *Manual for the Minnesota Satisfaction Questionnaire*. Minneapolis: University of Minnesota, Work Adjustment Project, Industrial Relations Center, 1967.

(Received November 20, 1972)

(*Continued from page 134*)

Effects of Spectral Differences in Illumination on Fatigue. James B. Mass (Center for Improvement of Undergraduate Education, Cornell University, Rand Hall, Ithaca, New York 14850), Jill K. Jayson, and Douglas A. Kleiber.

Stress and Performance During and After Officer Candidate School. J. Richard Jennings (Department of Experimental Psychophysiology, Walter Reed Army Institute of Research, Forest Glen Section, Building 189, Washington, D. C. 20012), R. M. Rose, and Leo Kreuz.

Comparisons of Self and Superior Ratings of Managerial Performance. Herbert G. Heneman III (Graduate School of Business, University of Wisconsin, 1155 Observatory Drive, Madison, Wisconsin 53706).

Role of Rater in Performance Appraisal. Richard J. Klimoski (Department of Psychology, College of Social and Behavioral Sciences, 404C West 17th Avenue, Columbus, Ohio 43210) and Manuel London.

Seat Belts: Relations between Beliefs, Attitude, and Use. Gunilla Fhanér (Kristina Hards vag 9, S-443 00, Lerum, Sweden) and Monica Hane.

Time Delay and Similarity Effects in Facial Recognition. Kenneth R. Laughery (Department of Psychology, University of Houston, Cullen Boulevard, Houston, Texas 77004), Paul K. Fessler, David R. Lenorovitz, and Darryl A. Yoblick.

Effect of Ambient Illumination on Noise Level of Groups. Mark S. Sanders (Department of Psychology, California State University, Northridge, California 91324), Jerry Gustanski, and Mike Lawton.

Journal of Applied Psychology
1974, Vol. 59, No. 2, 145–156

SOURCES OF SATISFACTION AND DISSATISFACTION AMONG SOLID WASTE MANAGEMENT EMPLOYEES[1]

EDWIN A. LOCKE[2] AND ROMAN J. WHITING

University of Maryland

A national random sample of 911 employees in the solid waste management industry were interviewed concerning their job attitudes. While white-collar employees were more satisfied with their jobs than were blue-collar employees, both the latter and the secretarial employees were more satisfied than were the employees at the same job levels in other types of work. These differences were attributed to pay and work differences between solid waste management and other occupations. White-collar employees were more likely to derive satisfaction and dissatisfaction from intrinsic sources and less likely to derive them from extrinsic sources than were blue-collar workers. Herzberg's motivator-hygiene theory of attitudes was not supported by the data.

Recent studies of blue-collar–white-collar differences in sources of job satisfaction and dissatisfaction have used three basic sampling methods: (*a*) population or random sampling of different job levels within the same plant (Armstrong, 1971; Gluskinos & Kestelman, 1971) or of different job levels at one location of a multiunit organization (Friedlander, 1965); (*b*) stratified random sampling of a cross section of employed persons in one urban area (Centers & Bugental, 1966) or in the United States as a whole (Gurin, Veroff, & Feld, 1960; Morse & Weiss, 1972); and (*c*) "accidental" sampling of employees in different occupations living in a single urban area (Locke, 1973).

Despite these different sampling procedures, the findings have been generally consistent with Herzberg, Mausner, Peterson, and Capwell's (1957) conclusions based on an integration of the earlier literature. White-collar workers are more likely to derive pleasure from intrinsic factors such as the work itself and work achievement, while blue-collar employees more often get pleasure from extrinsic factors such as physical or social working conditions and pay or security.

Minimal job level differences were found, however, by Myers (1964) at Texas Instruments, Dallas, who used Herzberg, Mausner and Snyderman's (1959) critical incident procedure and classification system. Herzberg's (1966) more recent work also deemphasizes individual, and therefore job level, differences in sources of satisfaction and dissatisfaction on the job. Schneider and Locke (1971) have criticized Herzberg's classification system on the grounds of logical inconsistency. Locke (1973), using the same critical incident method but a new classification system developed by Schneider and Locke (1971), obtained results consistent with the findings of Herzberg et al., 1957 with one accidental sample of white-collar and blue-collar employees, but not with another sample. Since there was some evidence to indicate that the latter failure to replicate was attributable to differences in occupational sampling, Locke recommended that future studies use employees from single occupational groups.

One purpose of the present study was to attempt to replicate earlier findings on white-collar–blue-collar differences using a nationwide random sample of employees from a *single* industry—solid waste management (i.e., collection and disposal of garbage and other solid waste). To date, according to the present investigators' knowledge, this is the only study of sources of satisfaction and dissatisfaction using a sample of this type. The present study used the Herzberg critical incident method, and the revised incident classification system (Locke, 1973), and included other questions as well.

---

[1] This study was supported by Contract No. 68-03-0041 between the Environmental Protection Agency (EPA) and Applied Management Sciences (AMS), Silver Spring, Maryland. The computer time for the analyses in this paper was supported in full through the facilities of the University of Maryland Computer Science Center.

The contents of this paper represent the views of the authors only and do not represent the official position of EPA or AMS; nor does the publication of this paper imply that the authors agree with the policies of EPA.

The authors wish to thank David Snyder and Todd Tucker of AMS, and John Hill of EPA for their help in designing, conducting, and analyzing this study.

[2] Requests for reprints should be sent to Edwin A. Locke, College of Business and Management, University of Maryland, College Park, Maryland 20742.

A second purpose of this study was to test Herzberg's motivation-hygiene theory using the revised incident classification system which calls for separate classification by events (what happened) and agents (who made it happen). Previous results using more restricted samples have shown consistently that motivator events were more likely to lead to both satisfaction and dissatisfaction than were hygiene events, but that the self was more likely to be given credit for satisfying events and others blamed for dissatisfying events rather than vice versa (Locke, 1973, in press; Schneider & Locke, 1971).

Due to the limited sample sizes, these previous studies did not report event-agent interactions on a category-by-category basis. The third purpose of the present study was to report such interactions for the first time. These data make it possible to answer such questions as the following: (*a*) With respect to events, which agents are seen as primarily responsible for each? (*b*) With respect to agents, which events are they most frequently seen as causing?

A fourth purpose was to compare solid waste employees to other occupational groups in terms of degree or amount of satisfaction. One item used in this study, and thought to be indicative of overall satisfaction, was also used in a number of previous studies with a variety of occupational groups (summarized in Sales & House, 1971). Therefore, it was possible to compare solid waste employees at various job levels to these groups. Since the job of garbage collector is almost universally considered to be the prototype of a "bad job" (e.g., low status, physically tiring, dirty, smelly, too hot, too cold, etc.), one might expect employees in this field to be less satisfied than employees in most other fields.

METHOD

The employee data were gathered as part of a larger nationwide study of the solid waste management industry. (A detailed description of the larger study is found in Applied Management Sciences, 1972). Management representatives from 3,327 public and private collection agencies were interviewed concerning company policies and manpower. In some of these firms employees were also interviewed concerning their attitudes and opinions.

Sampling Plan

The basic design involved a national stratified random sample of agencies engaged in solid waste management. The stratification variables were geographical region (7); city size (4) based on the Standard Metropolitan Statistical Area system with random selection of cities within each stratum (though in the largest two size categories there was 100% sampling); and public versus private with random sampling in each stratum for each city selected, though in the largest cities close to 100% of the public agencies were sampled.

All 3,327 agencies chosen by the previous method were asked to allow a management representative to submit to an interview concerning management policies. In addition, in approximately one out of every three or four agencies, the interviewer asked permission to conduct one employee interview. (In a few cases two or three employees were interviewed from the same agency.) The stratifying variables for the employee sample were geographical region; public versus private; and job level (5). Equal numbers of public and private employees were supposed to be interviewed within each region, since previous research had indicated the number of employees in the public and private sectors were approximately the same. (More recent data indicate there are more in the private sector.) The five job levels were: (*a*) managerial, (*b*) clerical/secretarial, (*c*) supervisory/foreman, (*d*) skilled labor, and (*e*) unskilled labor; the sample size was proportional to the estimated number of employees at each level in the population. Within each company, the employee was chosen at random from within the job category desired. (Sometimes the chosen employee was not available and the interviewer had to take pot luck.) The total number of employees interviewed at each job level is shown in the first row of Table 1.

Employee Questionnaire

The 46 questions asked of the employees fell into four main areas: (*a*) demographic information (age, education, etc.); (*b*) present job information (job title, tenure, hours of work, reason for taking present job, etc.); (*c*) rewards (previous raises, present pay, benefits, etc.); and (*d*) attitudes (amount of satisfaction, sources of satisfaction, etc.). Some information (name, job level, etc.) was obtained in advance from management.

The specific attitude items used for this report are described below.

Reason for working in solid waste field was determined from the following open-ended item: "How did you come to work in the solid waste management field? What made you decide to work in this field?"

Amount of satisfaction with each of 11 job aspects was measured by means of a 7-point "faces" scale with verbal anchors. The sum of the faces yielded a combined overall satisfaction measure.[3]

A single, indirect measure of satisfaction was derived from a yes or no response to the following item (used in previous studies; cf., Sales & House, 197[illegible]): "If you had your life to live over, would you like to wind up in the same line of work as the one you're doing now?"

Sources of satisfaction were determined from four items, two of the Herzberg type (one for satisfaction and one for dissatisfaction):

> Now would you try to think of a specific time when you were *satisfied* [dissatisfied] with your present job, a time when you felt especially *good* [bad] about your job. Can you think of such a time? Exactly what happened to make you feel good [bad]? What led up to this good [bad] feeling? Who or what was mainly responsible for this event?

and two of the conventional type:

> What *two* things about this job give you the *most pleasure* [displeasure]? What two things make you *happiest* [most unhappy] about your job here?

The order of asking for satisfiers and dissatisfiers was alternated by each interviewer.

Another open-ended item asked for the individual's work preoccupations:

> Now I would like you to think back to this morning [or any typical morning] when you were getting ready to come to work or were on your way to work. I would like to know what you were thinking about as you got ready to leave or as you were on your way. Can you remember this? . . . How did you feel? What kind of mood were you in?

Coding categories for the open-ended items are described in the Results section.[4]

Interviewers

All questionnaires were administered by interview. The interviewing staff consisted of 15 military retirees selected from a pool of over 400 on the basis of previous experience and other pertinent factors. All interviewees received from three to five full days of training, including: (*a*) an explanation of the study; (*b*) information concerning the solid waste industry; (*c*) an explanation of the employee questionnaire including the purpose of each item; (*d*) a chance to listen to tapes of previous interviews conducted during the pretest stage; and (*e*) practice interviews with simulated and/or actual employees followed by critiques from senior staff members. In addition, each interviewee was provided with a highly detailed interviewer's manual covering all aspects of the questionnaire and the interview process, and was quizzed as to its contents before being allowed to begin interviewing.

Employee interviews were conducted on-site but in complete privacy with a guarantee of total confidentiality. Interviews were verified by follow-up post cards mailed to the interviewees.

Editing and Coding

Completed interview forms were mailed back to the conducting agency weekly. Editors examined the questionnaires for errors (omissions, incomplete answers, contradictions etc.). Errors were corrected either by calling the employee directly (at home) or calling the interviewer. Interviewers were kept constantly informed of observed errors.

All open-ended items were coded independently by at least two coders. Disagreements, except in the case of the Herzberg-type questions, were resolved by discussion among the coders. The coding of the Herzberg-type questions was checked by a third coder (the first author was the third coder for the first half of the sample), since these were the most difficult items to code.[5] Records of interrater agreement could not be kept (due to production pressures on the coders), but generally disagreements occurred only when the interviewer himself had been incomplete or unclear in the answers he elicited and/or wrote. Often such ambiguities were clarified by call-backs.

RESULTS

Demographic and Pay Data

The second through fifth lines of Table 1 show pertinent demographic and pay data for the five job levels. Excluding the secretarial/clerical employees for the moment, it

---

[3] The 11 dimensions were: work itself, work variety, amount of work, pay, chances for promotion, supervision, co-workers, subordinates, hours of work, equipment, and benefits.

[4] All responses that were uncodable (according to the coding system used) were omitted from the analyses along with missing data caused by nonresponse to the question. Thus, the *ns* vary from table to table.

[5] The question of possible bias by the first author may be raised here. In answer it should be pointed out that in a previous study (Locke, 1973, Study 2) this author agreed with a naive rater on 84% of the event and 88% of the agent categorizations. Furthermore, this author was only one (and last) of three coders on half of the sample and did not look at the coding of the remaining half of the sample. None of the other coders were acquainted with the Herzberg theory.

TABLE 1

DEMOGRAPHIC AND OVERALL SATISFACTION DATA BY JOB LEVEL

Demographic data	Employee											
	Unskilled (n = 326)		Skilled (n = 303)		Supervisory (n = 146)		Secretarial/ clerical (n = 75)		Managerial (n = 61)		df	F
	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD		
Tenure (in years)	4.49	6.01	6.20	6.90	10.49	7.70	4.77	5.24	8.36	8.19	4/906	19.18*
Age (in years)	33.78	11.45	36.56	11.55	41.97	11.93	37.32	12.63	40.53	10.63	4/905	24.02*
Education (in years)	9.51	2.81	10.15	2.42	11.00	2.48	12.19	1.74	12.90	2.81	4/892	38.03*
Weekly gross pay	144.60	50.23	164.70	43.03	206.10	69.02	139.00	47.83	245.30	113.72	4/888	75.99*
Total satisfaction (sum of 11 faces scales)	58.56	10.91	60.19	9.25	64.90	7.69	65.84	8.83	64.45	8.90	4/857	21.50*
Indirect satisfaction (% yes) (Choose same line of work again?)	46.4		55.4		69.0		73.6		62.3		4	37.66^{a*}

Note. n for some means are less than the total figure given due to missing data.
[a] Chi-square; Yates's correction for continuity applied.
* $p < .001$.

is clear that the white collar employees are older, have more education and tenure, and are more highly paid than blue collar workers. All F ratios are significant at the .001 level. The pay and education differences require no explanation. The age and tenure differences can be attributed to the higher turnover rate among the lower level employees and the fact that garbage collection, being physically demanding, is a younger man's job. The secretarial/clerical employees more closely resemble the blue collar than the white collar employees with respect to tenure, age, and pay.

Amount of Satisfaction

The bottom two rows of Table 1 show the usual positive association between job level and job satisfaction for both measures (ps < .001). While pay differences are usually associated with job level differences by their very nature, it is relevant to ask whether age, education, or tenure could have been responsible for the satisfaction differences between job levels. To test this possibility, job level was dichotomized into blue collar–white collar (omitting the secretarial employees who did not consistently follow the pattern of the supervisors and managers). Job level and the variables in Table 1 were then intercorrelated across all employees. The results showed a significant ($p < .001$) relationship between job level and total satisfaction even when age, education, and tenure were partialled out.

The percentages in the bottom row are for the item: "If you had your life to live over, would you like to wind up in the same line of work as the one you're doing now?" Sales and House (1971) reported that in previous studies between 16% and 21% of unskilled steelworkers and automobile workers answered yes to this question, whereas 46% of unskilled solid waste employees gave the same answer. Similarly, between 41% and 52% of skilled printers, steelworkers, and automobile workers in previous studies gave a yes answer, while 55% of skilled solid waste employees said they would choose this same line of work again. Thus, at the blue-collar level, garbage collectors appear to be *more rather than less satisfied* with their jobs than workers at the same job level in other fields.

Sales and House reported no satisfaction data for supervisors but they noted that a nationwide study of clerical employees yielded 42% yes responses compared to 74% in the present sample. This difference in the favor of the solid waste group is even more striking than that for the blue-collar employees.

The managerial employees in this study gave 62% yes responses which is somewhat lower than the 69% figure in a nationwide sample of managers reported by Sales and House.

TABLE 2

TOTAL FREQUENCY DISTRIBUTION FOR SATISFIER AND DISSATISFIER EVENTS

Event	Satisfiers		Dissatisfiers		Total		χ^{2a}
	n	%	n	%	n	%	
Motivator							
Task activity	28	3.6	52	7.4	80	5.4	9.55*
Amount of work	34	4.3	107	15.2	141	9.5	45.77**
Smoothness	59	7.5	138	19.7	197	13.2	40.71**
Achievement	83	10.5	33	4.7	116	7.7	15.41**
Promotions	90	11.4	26	3.7	116	7.7	27.32**
Responsibility	33	4.2	10	1.4	43	2.9	8.78*
Recognition	124	15.7	96	13.7	220	14.8	0.92
Hygiene							
Money	185	23.4	61	8.7	246	16.5	48.11**
Interpersonal atmosphere	33	4.2	45	6.4	78	5.2	3.13
Working conditions	120	15.2	134	19.1	254	17.0	3.05
Totals	789		702		1,491		202.75**

[a] All chi-squares are based on 1 degree of freedom. Yates's correction applied.
* $p < .01$.
** $p < .001$.

Herzberg Theory and Event-Agent Interactions

The categories used for coding the Herzberg-type questions are the same as those used by Locke (1973) as adapted from Schneider and Locke (1971). Table 2 shows the results for events for all subjects combined on a category-by-category basis.

The overall chi-square was significant ($p < .001$) but not of direct relevance to the Herzberg theory. More pertinent are the frequencies for the individual categories. Since the rows of the table are not independent, the chi-squares are meaningful mainly as descriptive (rather than inferential) statistics.

Herzberg's theory (or Theory 3 according to King's (1970) classification of its five versions) predicts that motivators are more likely to act as satisfiers than as dissatisfiers, while the opposite holds true for hygienes. However, the data show that three of the seven motivator categories (task activity, amount of work, and smoothness of work) are more likely to act as dissatisfiers than as satisfiers. Three other motivators (achievement, promotions, and responsibility) show a pattern in agreement with the Herzberg theory, while one category (recognition) shows no significant trend. With respect to hygienes, money is more likely to act as a satisfier than as a dissatisfier, the opposite of Herzberg's theory, while interpersonal atmosphere and working conditions show no significant trend.

Thus, in 4 of the 10 comparisons, Herzberg's theory was contradicted; in 3 cases it was not supported; and in 3 cases it was supported.

The results for the combined motivator and hygiene categories are shown separately by job level in Table 3. The chi-square values for the 2 × 2 contingency tables are significant for the unskilled and skilled employees only. Furthermore, the relationships for these two subsamples are directly opposite to what the Herzberg theory predicts. In these groups, motivators are more likely to act as dissatisfiers and hygienes as satisfiers rather than vice versa. Generally, motivators were more likely than hygienes to cause both satisfaction and dissatisfaction in the employee groups.

The results for the individual agent categories are shown for the combined sample in Table 4. Again the chi-squares for each row are presented separately for descriptive purposes. The self, the supervisor, and the organization are more likely to be seen as causing satisfier than dissatisfier events, while the opposite is true for co-workers, subordinates,

TABLE 3

FREQUENCY OF MOTIVATORS AND HYGIENES AS SATISFIERS OR DISSATISFIERS BY JOB LEVEL

Employee	Satisfiers		Dissatisfiers		χ^{2a}
	n	%	*n*	%	
Unskilled					
Motivators	110	41	151	60	18.75**
Hygienes	161	59	101	40	
Skilled					
Motivators	143	53	159	65	6.57*
Hygienes	125	47	86	35	
Supervisory					
Motivators	100	76	78	74	.05
Hygienes	31	24	27	26	
Secretarial/clerical					
Motivators	51	78	42	72	.32
Hygienes	14	22	16	28	
Managerial, professional and technical					
Motivators	47	87	32	76	1.24
Hygienes	7	13	10	24	

[a] Yates's correction applied.
* $p < .05$.
** $p < .001$.

customers, and nonhuman agents (e.g., the weather).

Herzberg's theory fails to make any explicit distinction between agents and events, therefore, comparison with theoretical predictions is difficult. However, Herzberg does claim that supervision and organizational policies are primarily dissatisfiers, the opposite of the results obtained in this study.

There was a definite overall tendency for the self to be given credit for satisfying events and other (nonself) agents to be blamed for dissatisfying events ($\chi^2 = 13.93$, $p < .001$). However, this overall trend was not present at all among the unskilled employees.

The event-agent interactions are shown in Table 5 for those subjects for whom complete data were available. (Those indicating an event but no agent or an agent but no event were omitted, thus the *ns* in this table are smaller than those in Tables 2 and 4.) The satisfying and dissatisfying interactions are given in contiguous columns, with the otal for each pair. (The chi-square for the atisfier and dissatisfier interactions are highly significant, $ps < .001$.) While there are definite differences within some cells in the frequencies of satisfiers versus dissatisfiers (as expected, based on Tables 2 and 4), the most frequently mentioned satisfying event attributed to every agent was also the most frequently mentioned dissatisfying event attributed to that agent. For example, the self was most often mentioned as the cause of both achievement and failure.

The supervisor was most frequently mentioned as the source of both positive and negative recognition. The supervisor was also mentioned often as a source of promotions, money (e.g., raises), working conditions, and interpersonal relations. Co-workers were seen mainly as affecting interpersonal relations, while subordinates were most often perceived to affect the smoothness of the work.

The company or organization was viewed as causing the highest percentage of both satisfying and dissatisfying events, with emphasis on money, working conditions, promotions, amount of work, and recognition. Customers were seen mainly as agents of recognition, and unions as a source of higher pay (money). Typically the nonhuman agent was nature or the weather that caused unpleasant working conditions and made the work go less smoothly (e.g., as when it snowed or rained).

Overall the organization and the supervisor were seen as being by far the most frequent causes of both satisfying and dissatisfying events (with the exception of non-

TABLE 4

TOTAL FREQUENCY DISTRIBUTION FOR SATISFIER AND DISSATISFIER AGENTS

Agent	Satisfier		Dissatisfier		Total		χ^{2a}
	n	%	*n*	%	*n*	%	
Self	95	12.3	46	6.5	141	9.5	12.65***
Supervisor	212	27.5	122	17.2	334	22.6	16.96***
Co-workers	23	3.0	56	7.9	79	5.3	15.71***
Subordinates	14	1.8	35	4.9	49	3.3	9.89**
Organization	307	39.9	188	26.5	495	33.5	19.28***
Customers	49	6.4	71	10.0	120	8.1	5.64*
Nonhuman	57	7.4	176	24.8	233	15.8	70.00***
Union	13	1.7	15	2.1	28	1.9	.17
Totals	770		709		1,479		150.30***

[a] All chi-squares are based on 1 degree of freedom. Yates's correction applied.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

TABLE 5

FREQUENCY OF EVENT/AGENT INTERACTIONS FOR SATISFYING AND DISSATISFYING INCIDENTS (ALL SUBJECTS)

Event	Agent																												
	Self			Supervisor			Co-worker(s)			Subordinate(s)			Organization			Customer(s)			Nonhuman			Union			Total		Grand total
	S^a	D^b	Total^c	S^a	D^b	Total	S^a	D^b	Total	S^a	D^b	Total	S^a	D^b	Total	S^a	D^b	Total	S^a	D^b	Total	S^a	D^b	Total	S^a	D^b	
Task activity	8	2	10	7	14	21		1	1		2	2	13	15	28		6	6		5	5		1	1	28	46	74
Amount of work		2	2	5	12	17		13	13	1	3	4	13	30	43	2	15	17	2	14	16		1	1	23	90	113
Smoothness	7	4	11	2	10	12	4	11	15	6	19	25	3	10	13	2	8	10	18	70	88			—	42	132	174
Achievement	57	20	77	6	2	8	1		1	5	2	7	7	4	11	2	1	3	1	2	3			—	79	31	110
Promotion	7	3	10	35	5	40	1		1			—	44	16	60			—			—	1	1	2	88	25	113
Responsibility	2	1	3	15	3	18	1	1	2			—	15	5	20			—			—			—	33	10	43
Recognition	2	1	3	60	37	97	6	4	10		1	1	20	17	37	33	35	68	1		1			—	122	95	217
Money	4	2	6	31	9	40			—			—	135	45	180	5		5		2	2	9	2	11	184	60	244
Interpersonal atmosphere		1	1	16	16	32	9	16	25		3	3	5	4	9	3	1	4		1	1		1	1	33	43	76
Working conditions	6		6	24	10	34	1	4	5	2		2	37	29	66	1	5	6	34	79	113			—	105	127	232
Total	93	36	129	201	118	319	23	50	73	14	30	44	292	175	467	48	71	119	56	173	229	10	6	16	737	659	1,396^d

[a] Satisfying.
[b] Dissatisfying.
[c] Total of Satisfying plus Dissatisfying.
[d] Agent-event data omitted for subjects indicating event but no agent or agent but no event.

TABLE 6

SATISFIER AND DISSATISFIER EVENT FREQUENCIES FOR BLUE-COLLAR AND WHITE-COLLAR EMPLOYEES

Event	Satisfier					Dissatisfier				
	Blue-collar		White-collar[a]		χ^{2b}	Blue-collar		White-collar[a]		χ^{2b}
	n	%	*n*	%		*n*	%	*n*	%	
Task activity	17	3.2	8	4.3	.25	35	7.0	13	8.8	.27
Amount of work	31	5.8	2	1.1	5.56*	89	17.9	10	6.8	8.39**
Smoothness	30	5.6	20	10.8	4.71*	75	15.1	48	32.7	17.36***
Achievement	33	6.1	41	22.2	33.15***	16	3.2	12	8.2	5.27*
Promotions	43	8.0	45	24.3	28.90***	18	3.6	7	4.8	.15
Responsibility	14	2.6	10	5.4	2.54	6	1.2	3	2.0	.10
Recognition	85	15.8	21	11.4	1.55	71	14.3	17	11.6	.44
Money	148	27.5	29	15.7	7.32**	48	9.7	7	4.8	2.68
Interpersonal atmosphere	28	5.2	2	1.1	4.72*	26	5.2	12	8.2	1.17
Working conditions	110	20.4	7	3.8	22.54***	113	22.7	18	12.2	5.63*
Totals	539		185		111.24***	497		147		11.46***

[a] Secretarial and clerical employees omitted.

[b] All chi-squares are based on 1 degree of freedom. Yates's correction applied.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

human, which played a major role in causing dissatisfying events). The company, for example, was seen as the major causal agent in 5 of the 10 event categories (task, amount of work, promotion, responsibility, and money), while the supervisor was the most frequently mentioned agent for two others (recognition and interpersonal atmosphere).

Blue-Collar–White-Collar Differences in Sources of Satisfaction and Dissatisfaction

For the purpose of making blue-collar–white-collar comparisons, the secretarial/clerical employees were deleted. The main reason was that a substantial proportion of this group was female. Since employees in the remaining job categories were all male, this would have led to a confounding of sex with job level when making comparisons. The blue-collar category, then, consisted of the unskilled and skilled employees while the white-collar group was composed of supervisory and managerial (professional, etc.) personnel. It should be noted that in this study the terms job level and white-collar–blue-collar are used interchangeably. They are not viewed as separate concepts but as alternative ways of referring to the same basic distinction.

Table 6 shows the Herzberg event category results by satisfier versus dissatisfier and blue- versus white-collar. The column of chi-squares were again calculated for descriptive purposes. Considering first the satisfier events, blue-collar employees were more likely to cite amount of work, money, interpersonal atmosphere, and working conditions as satisfiers, while white-collar employees were more likely to derive satisfaction from smoothness of work, achievement, and promotions. The pattern for dissatisfaction was parallel, with blue-collar employees more often being dissatisfied by amount of work and working conditions, and white-collar employees being more frequently dissatisfied by (the lack of) smoothness and (the lack of) achievement in the work.

The results for agents shown in Table 7 reveal that white-collar employees were more prone to see the self and subordinates as causing satisfaction, while blue-collar workers were more likely to see the organization as a source of satisfaction. White-collar employees more frequently attributed dissatisfying events to subordinates and the union, while blue-collar workers more often saw their supervisors and co-workers as agents of dissatisfaction.

TABLE 7

SATISFIER AND DISSATISFIER AGENT FREQUENCIES FOR BLUE AND WHITE-COLLAR EMPLOYEES

Agent	Satisfier					Dissatisfier				
	Blue-collar		White-collar[a]		$\chi^{2\text{b}}$	Blue-collar		White-collar[a]		$\chi^{2\text{b}}$
	n	%	*n*	%		*n*	%	*n*	%	
Self	42	8.0	39	21.2	19.37**	30	6.1	9	5.7	.001
Supervisor	136	26.1	53	28.8	0.29	100	20.2	17	10.7	5.60*
Co-workers	15	2.9	3	1.6	0.41	45	9.1	6	3.8	3.73
Subordinates	3	0.6	11	6.0	17.39**	7	1.4	27	17.0	53.00**
Organization	231	44.3	59	32.1	4.63*	129	26.1	41	25.8	.01
Customers	39	7.5	8	4.3	1.55	52	10.5	9	5.7	2.55
Nonhuman	44	8.4	10	5.4	1.23	126	25.5	40	25.2	.01
Union	12	2.3	1	0.5	1.43	5	1.0	10	6.3	12.39**
Totals	522		184		46.30**	494		159		77.29**

[a] Secretarial and clerical employees omitted.
[b] All chi-squares are based on 1 degree of freedom. Yates' correction applied.
* $p < .05$.
** $p < .001$.

Since each employee was instructed to make two responses to each sources of pleasure and displeasure item, the two responses were combined for purposes of analysis. (In some cases, the employee only made one response.) Since the employees were not asked specifically to cite agents or events, the categorizing system was necessarily based on the answers they gave spontaneously. The coding categories turned out to be similar to the revised Herzberg categories used in the previous analyses except that events and agents were not kept separate.

The results (not shown) were that white-collar employees mentioned the work itself, achievement, smoothness, subordinates, and customers as sources of pleasure more frequently than did blue-collar employees. The latter more often derived pleasure from the amount of work, pay (benefits/security), co-workers, and working conditions. Blue-collar employees were also more likely to derive displeasure from these same sources than were white-collar employees. The latter were more often displeased by subordinates and customers.

The above pattern of results was further verified by the employees' descriptions of (*a*) their thoughts or preoccupations while coming to work in the morning and (*b*) of their reasons for working in the solid waste field. Blue-collar employees were more preoccupied with being on time, the amount of work, and working conditions than were white-collar workers. They were also more likely to be thinking of quitting and of off-the-job activities, presumably due to their greater dissatisfaction with their jobs as compared to white-collar employees (see Table 1). Blue-collar employees more often took their jobs mainly because they needed a job. In contrast, the white-collar employees were more preoccupied with the content of the work to be done and with subordinates and were more likely to have taken their job because they like the type of work.[6]

DISCUSSION

The previously obtained differences between white-collar and blue-collar employees with respect to sources of satisfaction was replicated for the first time with a nationwide sample of workers in the same occupation. The intrinsic–extrinsic dichotomy affords a reasonably accurate way of summarizing the nature of these differences. The white-collar employees were relatively more often affected by the work itself, achievement, smoothness

---

[6] Tables showing the actual category frequencies and relevant descriptive statistics for the items described in the latter three paragraphs are available from the authors upon request.

(a kind of short-term achievement), by rewards that are associated with changes in work content (promotions), and by the agents most often associated with the foregoing events (self in the case of achievement, and subordinates in the case of smoothness).

In contrast, blue-collar employees were more often satisfied or dissatisfied by extrinsic rewards such as money and benefits, by work context factors such as working conditions, and interpersonal relations, and by co-workers who were associated with the latter type of event. The one exception to this trend was that blue-collar workers were more affected than white-collar employees by the amount of work. While amount of work is not extrinsic to the work itself, it is separate from work content, work progress (smoothness), and work outcome (achievement). The blue-collar orientation in this sample appears to involve a view of work as a means to an end rather than an end in itself; the work itself is liked to the extent that there is not very much of it. (Blue-collar employees rarely complained about too little work.) Presumably avoidance of fatigue is a major goal of employees with physically demanding work. Chadwick-Jones (1969) found reduced physical effort to be a major source of satisfaction for English workers who had been transferred from a preautomated to an automated steel plant.

These white-collar–blue-collar differences should not necessarily be interpreted as indicating that the two groups of employees have (on the average) different value systems. It is equally plausible to assume that the two groups have simply had different work experiences. For example, in the present study, of the 455 satisfying events reported by employees whom the company claimed had never received a promotion, only 10 (2%) described a promotion as a satisfying event. (Presumably these 10 individuals defined promotion differently than their bosses or else the company records were incorrect.) In contrast, of the 334 satisfying events reported by employees who had received a promotion, 80 (24%) described a promotion as a satisfying incident. To take another example, white-collar employees were much less likely than were blue-collar employees to mention the weather as a source of satisfaction or dissatisfaction. One reason may have been that the former do their work indoors while most of the latter work outdoors. Similarly, blue-collar employees did not often mention subordinates as sources of satisfaction or dissatisfaction (Table 7) because most of them did not have any subordinates.

To restate the obvious, an event must actually happen before it can be described as a source of satisfaction or dissatisfaction. If an event is not mentioned often, it could mean either that it did not occur or that it occurred but the individual did not appraise it as significant enough to be worth reporting.

With respect to Herzberg's theory, little support for its main tenets was found at any job level in this study. Previous studies using the same classification system employed in this study found that motivators were the major source of both satisfaction and dissatisfaction. The same result was obtained in this study, except that the trend was more marked at the higher job levels.

The major event-agent interactions are most easily interpreted from an instrumental or functional point of view. The agents can be viewed as facilitating or blocking employee value attainment (Locke, 1970, in press). These values include task values (content, amount of work, smoothness, achievement), valued rewards for task performance (responsibility, promotions, money, and recognition), and conditions which are congruent with physical needs and which facilitate goal attainment (working conditions). The supervisor and the organization were found to have the most pervasive influence as both value facilitators and value blockers, whereas the other human agents were seen as instrumental in attaining only particular values, for example, the self and achievement, subordinates and smoothness, customers and recognition, and unions and money.

Interpersonal relations, associated mainly with co-workers and supervisors, seem to entail valuing the agent as an end in himself rather than for his functional value. This type of relationship has been described by Locke (in press) as an entity relationship to distinguish it from the functional relationships of the type previously described. Entity relationships did not prove to be of great importance to the employees in this sample.

A final issue concerns the reasons for the greater satisfaction among the unskilled and skilled employees in this study as compared with those at the same job level in different occupations. One could begin by asking whether the item used to measure satisfaction actually measures it. This item correlated .37 ($p < .001$) with the total of the facets scales for all subjects combined. While this correlation was not very high, the single item suffered from restriction of range and might not be highly reliable for a given individual; however, this should not be the case when groups are compared. Also, the item was the same one used in previous studies. What then accounts for the different results?

With respect to pay, the gross wages of solid waste employees were somewhat higher than those of employees in other industries. The gross weekly base pay of nonsupervisory, privately employed, nonagricultural workers in the United States in 1971 was $125 (Statistical Abstract, 1971) as compared to a mean figure of $128 for unskilled and skilled solid waste employees combined (based on data obtained from management in the present study; see Applied Management Sciences, 1972). The latter figure does not include overtime pay which raised the mean weekly gross pay of the blue-collar employees in this sample to over $150. There were also other respects in which the unskilled and skilled solid waste employees may have advantages over other occupations. The unskilled workers are typically collectors while many of the skilled workers are truck drivers, who may help with collection. (The national average is approximately two men per truck.) The fact of being alone with co-workers on a truck means considerable autonomy in terms of freedom from supervision. The fact that a different neighborhood is covered each day means that there is some degree of variety in the work task. In addition, the type of solid waste collected may differ day to day, from season to season, and from location to location. Since the blue-collar workers are in charge of their own truck, there is considerable freedom to control pace of work and even freedom to choose methods of work (e.g., where to stop the truck, how to carry the waste containers, how to coordinate teamwork). All of these characteristics have been found to be associated with high work satisfaction in previous studies (for a summary see Locke, in press). In these respects the job of solid waste collector differs markedly from the typical assembly line factory job where close supervision, repetitiveness, and little freedom of choice with respect to either pace or method of work are characteristic.

Finally, while the following did not show up in the figures of this study, there is reason to believe that solid waste collectors have shorter working hours than most other skilled and unskilled workers. Many solid waste collectors, especially those in the public sector, are on the task system, whereby they may leave whenever they complete their assigned route for the day (Applied Management Sciences, 1972). This route sometimes requires only 5–6 hours work to complete. The reasons that these employees did not actually report working such short hours (mean reported hours were 43) may be that they are officially paid for a 40-hour week plus overtime.

While the above factors may help to explain the greater satisfaction of the blue-collar employees as compared to those in other professions, none of these considerations apply to secretarial/clerical employees. The high degree of apparent satisfaction shown by this group may be partially explained by their high pay. The reported weekly gross pay of the secretarial/clerical employees in this sample was $114, whereas the national median (mean was not available) for female employees in this category was $80 in 1969 (Statistical Abstract, 1971). Even adding 10% to this to compensate for the two-year difference (the present data were collected during 1971 and early 1972) yields a remaining difference of $26 per week. This means that solid waste employees in the secretarial/clerical category are paid 29% higher than the national average.

REFERENCES

APPLIED MANAGEMENT SCIENCES. Solid waste management manpower: Profile and analysis. (G-7, Final Report) Silver Spring, Md.: 1972.

ARMSTRONG, T. B. Job content and context factors related to satisfaction for different occupational levels. *Journal of Applied Psychology*, 1971, **55**, 57–65.

CENTERS, R., & BUGENTAL, D. Intrinsic and extrinsic job motivations among different segments of the working population. *Journal of Applied Psychology*, 1966, **50**, 193–197.

CHADWICK-JONES, J. K. *Automation and behaviour.* New York: Wiley, 1969.

FRIEDLANDER, F. Comparative work value systems. *Personnel Psychology*, 1965, **18**, 1–20.

GLUSKINOS, U. M., & KESTELMAN, B. J. Management and labor leaders' perception of worker needs as compared with self-reported needs. *Personnel Psychology*, 1971, **24**, 239–246.

GURIN, G., VEROFF, J., & FELD, S. *Americans view their mental health.* New York: Basic Books, 1960.

HERZBERG, F. *Work and the nature of man.* New York: World Publishing, 1966.

HERZBERG, F., MAUSNER, B., PETERSON, R. O., & CAPWELL, D. F. *Job attitudes: Review of research and opinion.* Pittsburgh, Pa.: Psychological Service of Pittsburgh, 1957.

HERZBERG, F., MAUSNER, B., & SNYDERMAN, B. *The motivation to work.* New York: Wiley, 1959.

KING, N. Clarification and evaluation of the two-factor theory of job satisfaction. *Psychological Bulletin*, 1970, **74**, 18–31.

LOCKE, E. A. The supervisor as "motivator": His influence on employee performance and satisfaction. In B. M. Bass, R. Cooper, and J. A. Haas (Eds.), *Managing for accomplishment.* Lexington, Mass.: Heath Lexington, 1970.

LOCKE, E. A. Satisfiers and dissatisfiers among white-collar and blue-collar employees. *Journal of Applied Psychology*, 1973, **58**, 67–76.

LOCKE, E. A. The nature and consequences of job satisfaction. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology.* Rand-McNally, in press.

MORSE, N. C., & WEISS, R. S. The function and meaning of work and the job. In V. H. Vroom & E. L. Deci (Eds.), *Management and motivation.* Baltimore: Penguin Books, 1972.

MYERS, M. S. Who are your motivated workers? *Harvard Business Review*, 1964, **42**(1), 73–88.

SALES, S. M., & HOUSE, J. Job dissatisfaction as a possible risk factor in coronary heart disease. *Journal of Chronic Diseases*, 1971, **23**, 861–873.

SCHNEIDER, J., & LOCKE, E. A. A critique of Herzberg's incident classification system and a suggested revision. *Organizational Behavior and Human Performance*, 1971, **6**, 441–457.

Statistical Abstract of the United States, (92nd ed.), Washington, D.C.: U.S. Bureau of the Census, 1971.

(Received December 20, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 2, 157–162

EFFECTS OF APPLICANT STEREOTYPES, ORDER, AND INFORMATION ON INTERVIEW IMPRESSIONS[1]

MANUEL LONDON[2] AND MILTON D. HAKEL

Ohio State University

Carlson's study of the effects of favorability, order, and type of information on interviewers' decisions was replicated. The results showed that unfavorable information is more important to interviewers' ratings than favorable information. However, this effect is moderated by the expected typical applicant—an expected typical applicant of high caliber leading to a more favorable first impression of unfavorable information. Unfavorable information showed contrast effects moderated by the type of information. Interviewers' confidence in evaluations was greater for favorable than for unfavorable information.

Since Webster's (1964) publication, numerous factors influencing interviewer's decisions have been identified. Broadly viewed, these many studies constitute two streams of research: (*a*) the impact of information characteristics including ordering of information (e.g., Carlson, 1971; Hakel, Ohnesorge, & Dunnette, 1970; Hollmann, 1972; Rowe, 1963; Springbett & Crowell cited in Webster, 1964; Wexley, Yukl, Kovacs, & Sanders, 1972), and the roles of stereotypes (e.g., Hakel, 1971; Hakel, Hollmann, & Dunnette, 1970; Mayfield & Carlson, 1966; Springbett, Sydiaha, & Rowe cited in Webster, 1964). This article reports a study which combines both streams, and also illustrates the impact of yet another source of variance in judgments, the interviewer's conception of the expected typical applicant.

It is clear that stereotypes have an influence on interviewers' decisions. Sydiaha (Webster, 1964) found that stereotypes used by an interviewer are not unique to the individual but are common to all who have a reasonable amount of experience in a specific employment situation. Rowe (Webster, 1964) found that the interviewer has a stereotype of the "good" rather than the "poor" worker and Sydiaha found that they all tended to look at the same information. A possible explanation for this generality of results is that the interviewers used as subjects were all basically the same type of person in each study with similar experiences and ideas as to what type of person is best for the job and what type of person tends to apply. (The interviewers, in fact, were all officers in the Canadian Army.) Mayfield and Carlson (1966) suggested that the stereotype of the ideal applicant is composed of two parts: one part dealing with those favorable and unfavorable individual characteristics on which there is interinterviewer agreement and the second part dealing with specific characteristics for individual interviewers. It is likely that interviewers from widely divergent backgrounds having had different experiences might have different stereotypes and might arrive at different decisions.

Interviewers in divergent situations are exposed to differing populations of applicants, and even if the applicants come from the same population (as within a local labor market at a single company location), interviewers see different samples of applicants. These population and sample differences may be manifest in interviewers' perceptions as differences in expectations about the overall level of qualification of the typical applicant. The expected typical applicant, similar to the stereotype of the ideal applicant in that both are abstracted from experience, can be conceptualized as the interviewer's estimate of mean level of overall qualification of a sample of applicants. Thus, while interviewers may share the same ideal applicant stereotype, their differing experiences may have produced differing expectations about the typical applicant. The expected typical applicant could function as an anchor, or perhaps even the adaptation level, on an interviewer's overall evaluation continuum, and judge to judge differences in the location of that anchor could be related to interjudge disagreement. This study investigates both ideal applicant stereotypes and expectations of the typical applicant as determiners of interview impressions, using Carlson's (1971) methodology as a starting point for replication and extension.

---

[1] This research was supported by Grant GS-2355 from the National Science Foundation.

[2] Requests for reprints should be sent to Manuel London, Department of Psychology, Ohio State University, Columbus, Ohio 43210.

Carlson (1971) tried to determine to what extent an initial valid impression established by knowledge of a test score is altered by subsequent information differing in favorability, order of presentation, and quality of preceding applicant. He found that information which is consistent with an interviewer's initial impression added less to his total impression than information of opposite value detracted from it. Further, the effect of the same unfavorable information was greater after a positive impression had been established than was the effect of positive information after a negative impression had been established.

Carlson's results are in line with Springbett's finding, reported by Webster (1964), that interviewers are more influenced by unfavorable than by favorable information. It seems that the interview is primarily a search for negative evidence, since the unfavorable impression was followed by rejection in 90% of the cases investigated. Rowe (Webster, 1964) found similar results with unfavorable and favorable typewritten descriptions of candidates.

This experiment examines the joint impact of several independent variables: experimentally induced *ideal* applicant stereotypes, experimentally induced typical applicant expectations, order of information, favorableness of information, and type of information. The type of information evaluated was subjective self-descriptions given by the ratees versus a valid selection test score given at the end of the interview.

METHOD

Subjects

Webster (1964) commented with some exasperation on the difficulties of getting experienced interviewers to adopt a common frame of reference. Therefore, eighty inexperienced college students enrolled in introductory psychology courses and an upper level psychology course at the Ohio State University served as raters. No control was made for sex.

Experimental Design

Four induction combinations were used for studying the impacts of ideal applicants and typical applicants: (*a*) high qualified ideal applicant, high qualified expected typical applicant (HH); (*b*) high qualified ideal applicant, low qualified typical applicant (HL); (*c*) low qualified ideal applicant, high qualified typical applicant (LH); and (*d*) low qualified ideal applicant, low qualified typical applicant (LL). Within each of these combinations there were four types of information presentation describing two applicants. The sequences of favorable and unfavorable information were chosen to allow for examination of relevant order effects. Each applicant consisted of two paragraph descriptions and a test score described as a valid achievement test (the test score was given last so as not to confound the initial effect of the stereotypes). The favorability of the test score corresponded to the favorability of the first paragraph of information for the first applicant presented while the favorability of the test score for the second applicant corresponded to the favorability of the second paragraph of information. The possible sequences of favorability of each description for the two applicants presented to each subject were as follows: fff, uuu; fuf, uff; ufu, fuu; uuu, fff (u and f indicating favorability).

Materials

Descriptions of the ideal person for the job and the type of employee to expect were given to the interviewers. As a manipulation check and in an effort to get the interviewers to concentrate on and "internalize" these descriptions, they were asked to evaluate them on the same response scales as they would later use in the experiment. A 2 × 2 analysis of variance performed on this data indicated significant differences in favorability among the ideal applicant and the expected typical applicant descriptions.

The favorable and unfavorable items of information consisted of hypothetical self-descriptions of an employee's previous morning on his present job. The employees' descriptions were selected on the basis of ratings by 33 judges drawn from the target population. Ratings were made on the scales used in the main study. The descriptions were selected for use in the experiment according to the following criteria: (*a*) Their respective scale values differed significantly from the neutral point (0) but in opposite directions from it. (*b*) Favorable and unfavorable descriptions did not differ significantly in terms of absolute distance from the neutral point. (*c*) The descriptions conceivably could have referred to the same individual (Briscoe, Woodyard, & Shaw, 1967). The descriptions were systematically randomized for the prescale measure such that no subject received and rated the descriptions in the same order.

The response sheets contained thirteen 7-point scales of the interviewers' impressions of how well the employee met the job requirements listed on the scales. The scores ranged from 1 (very poor) to 7 (very good) and a sum of all 13 scores was taken as the total rating. An overall rating and the confidence the rater had in his decisions were also taken for each description, each on a 7-point scale.

Procedure

The data were collected during class meetings. Booklets of descriptions and scales were handed out so that subjects were randomly assigned to the experimental conditions. Descriptions of the ideal applicant and expected typical applicant were presented and rated first. The subjects were then asked to rate each applicant three times, once after each set of information was presented. No time limit was placed on completion of the booklet. Each subject rated two applicants. Subjects were reminded of the stereotypes before rating the second applicant.

One problem with the interpolated method of rating is that feeding information to a judge piece by piece affects the decisions (Webster, 1964). A judgment is both different and better when information is given to a judge simultaneously than when it is fed to him piece by piece. The studies cited by Webster indicate that it makes no difference in the final decision if the judge recorded preliminary impressions or merely made a mental note of each additional bit of information. Briscoe et al. (1967), however, found that a recency effect resulted from the interpolated mode of responding whereas a primacy effect resulted from the final mode. Carlson (1971) found that recency and primacy effects occurred in an interpolated mode of decision making for unfavorable information but not for favorable information. Briscoe et al., in fact, also found that to be true of unfavorable information when their data were examined for the order effects resulting from the favorability of the information. Thus, the interpolated mode of response does not inevitably result in recency effects.

Analysis

The design of the experimental material was a $2 \times 2 \times 4 \times 2 \times 3$ factorial. The independent variables were the ideal applicant stereotype, the expected typical applicant, the four orders of information favorability combinations, the position (first or second) of the two applicants rated by each subject, and the three paragraphs of information for each applicant rated. Factorial analyses of variance with repeated measures on the last two factors were performed on the treatment combinations for the total, overall, and confidence ratings.

RESULTS

Means and standard deviations were computed for the totals of the individual ratings, the overall ratings, and the confidence ratings. The Pearson product-moment correlation for the total score and the overall score was .88, and therefore only the results for the total scores are presented (see London, 1972, for means, standard deviations, and other results).

The results of the analysis of variance on the totals scores are given in Table 1 which includes the values of omegas squared (W^2), the proportion of variance accounted for by the independent variable (Hays, 1963).

The analysis of variance on the totals of raters' evaluations for each description showed that the interaction of order of information favorability with applicant position (first or second) accounted for the largest amount of variance: 45%. Differing combinations of paragraphs and order of information favorability accounted for 11% of the variance. Levels of ideal and typical applicants in combination with differing paragraphs and orders of information favorability also showed significance but accounted for a much smaller percentage of variance. The significant effect of the ideal applicant conditions indicates that the ratings were not made merely on a basis of each description as in the prescale ratings.

Several post hoc hypotheses were developed and tested for statistical significance. The hypotheses were as follows:

1. For hypothetical applicants consisting of consistent information—that is, all favorable or all unfavorable: (*a*) The difference between the first two favorable descriptions is not statistically greater nor less than the discrepancy between the second and third descriptions for consistently favorable applicants. (*b*) The difference between the first two unfavorable descriptions for a consistently unfavorable applicant is statistically greater than the discrepancy between the second and third unfavorable descriptions.

2. For hypothetical applicants constructed of inconsistent information—that is, fuf, uff, ufu, and fuu: (*a*) Unfavorable information has a statistically greater effect on a rater's evaluations when presented after rather than before favorable information when summed over all ideal applicant and typical applicant conditions. (*b*) A comparison between the

TABLE 1

ANALYSIS OF VARIANCE FOR TOTAL SCORES FROM THE 13 RATING SCALES

Source	Error	*df*	MS	*F*	W²
Ideal applicant (A)	F	1	2851.88	6.60*	.01
Expected typical applicant (B)	F	1	795.68	1.84	.00
Favorability combinations (C)	F	3	1097.77	2.54	.10
Position (D)	D × F	1	53.33	.13	.00
Paragraphs of information (E)	E × F	2	123.73	1.74	.00
A × B	F	1	26.13	.06	.00
A × C	F	3	490.38	1.13	.00
B × C	F	3	1035.78	2.40	.01
A × D	D × F	1	1235.21	2.92*	.00
B × D	D × F	1	2.41	.01	.00
C × D	D × F	3	35249.41	83.26**	.45
A × E	E × F	4	63.53	.89	.00
B × E	E × F	4	187.05	2.63*	.00
C × E	E × F	12	2034.21	28.63**	.11
A × B × C	F	3	499.24	1.15	.00
A × B × D	D × F	1	403.33	.95	.00
A × C × D	D × F	3	1660.42	3.92*	.02
B × C × D	D × F	3	288.79	.68	.00
Subjects (F)		64	432.36		
A × B × E	E × F	4	69.55	.98	.00
A × C × E	E × F	12	47.09	.66	.00
B × C × E	E × F	12	207.50	2.92*	.01
A × B × C × D	D × F	3	595.23	1.41	.00
D × F		64	423.37		
A × B × C × E	E × F	12	58.24	.82	.00
E × F		256	71.06		

Note. E is nested within D and F is nested within A, B, and C.
* $p < .05$.
** $p < .01$.

means of the first two ratings under stereotype combinations high qualified ideal applicant, high qualified typical applicant and low qualified ideal applicant, high qualified typical applicant are greater than the means of the first two ratings under the stereotype combinations high qualified ideal applicant, low qualified typical applicant, and low qualified ideal applicant, low qualified typical applicant. (This should also be true for the final ratings, had the stereotype treatment effect held up, but the results did not indicate this. The results also showed that the effect of unfavorable information was even less when the ideal applicant stereotype as well as the expected typical applicant were high. Favorable information, however, did not seem to be altered by either induction.)

These hypotheses were tested by the Scheffé method of orthogonal contrasts (Hays, 1963). Except for hypothesis 1(*a*), all were significant at the .05 level. Thus, unfavorable information does have a greater effect than favorable information though the effect of unfavorable information is modified by both the ideal applicant and expected typical applicant conditions. Unfavorable information was rated more favorably in the high qualified ideal applicant, high qualified typical applicant condition than in the low qualified ideal applicant, high qualified typical applicant condition, but was rated even lower in the high qualified ideal applicant, low qualified typical applicant condition, indicating that the expected typical applicant is the crucial variable in influencing the initial impression of the applicant.

DISCUSSION

The original analyses and the post hoc analyses yielded the following conclusions:

1. An interviewer who expects to find a high caliber applicant gives applicants higher ratings. This was shown by the significant dif-

ferences among interviewer judgments for the ideal applicant conditions and by the significant interaction of expected typical applicant with favorability combinations and paragraphs. The post hoc analyses, however, showed that it is the unfavorable information which diminishes in effect when an interviewer has the concept of a high caliber typical applicant and even more so for a high ideal applicant also, but the effect of favorable information remains unchanged by stereotype condition. The moderating variable in this study seems to be the expected typical applicant, a high typical applicant leading to more favorable ratings of unfavorable information. It does not seem, as Carlson (1971) found, that the rater overreacts to information of opposite view to his initial impression (or in this case the induced stereotype). But this finding does not hold for a low expected typical applicant making favorable information any less favorable than when it is preceded by a high typical applicant. Initial favorable information is rated about equally no matter what the stereotype, and a candidate who is overqualified, tends to be rated highly regardless of stereotype.

2. Consistently unfavorable information has stronger effects than consistently favorable information as shown by the results of the post hoc analysis 1(*b*). That is, the second paragraph of unfavorable information was rated more unfavorably than the first for an applicant comprised of consistently unfavorable information following an applicant of consistently favorable information. This hypothesis was not supported in the present experiment for the uuu order when it described the first applicant rated with no applicant of favorable quality preceding it. Analysis of the first bit of information in this sequence for the high qualified ideal applicant, high qualified typical applicant stereotype condition indicates that unfavorable information has a diminished effect after this stereotype is induced. It may be, then, that the interviewer is more unsure of his ratings on the first unfavorable description after rating the consistently favorable applicant or after having been induced to expect a high caliber applicant and therefore tends to decrease the ratings made after the second, but leveling off for the third and final rating.

3. Unfavorable information results in recency and sandwich effects. That is, fu results in a more unfavorable final rating than a uf sequence when a favorable test score is given last. That is not the case, however, when the same ratings are compared for order three, the final u of the ufu order applicant being no less unfavorable than the final u of the fuu applicant. It may be, then, that when an unfavorable test score is presented to an interviewer, the maxim is, as Carlson (1971) suggests of his failing test scores serving as initially valid information, "When the selection test is failed, forget the man, and when the selection test is passed, forget the test and evaluate the applicant using other tools."

4. The effects of induced stereotypes diminish or are altered by subjective information as the rater evaluates the applicant descriptions. Such distinctions as made above about the effects of stereotypes on first impressions cannot be made for final ratings for the second applicants since no such trends appear in the data even though the interviewers were asked to review the stereotypes before they read the first description of the second applicant.

5. Finally, the analysis of variance for confidence ratings (reported in London, 1972) and post hoc analyses on these data showed that interviewers seem to be more confident in making ratings of consistently favorable descriptions. This result is in line with the finding that while favorable information tends to be rated at the same level of favorability regardless of stereotype induction, the effects of unfavorable information seem to be moderated by a high expected typical applicant. Thus, interviewers may be less sure of their unfavorable ratings but believe that favorable applicants can succeed on the job no matter what the induced stereotype.

Since the type of methodology used in this study is commonly used in studies of impression formation (cf., Hastorf, Schneider, & Polefka, 1970, pp. 49–52) and investigations of decision making in the employment interview (e.g., Carlson, 1971; Wexley et al., 1972), it is useful to look at several problems with this type of method.

One problem is in generalizing results to the selection interview from studies of impression formation using a paper and pencil instrument with college students as subjects. Carlson (1971) tried to counteract this problem by using life insurance agency managers who actually do the hiring and are familiar with what the best man for the job should be like and the type of person who generally applies. The present experiment sought to manipulate the stereotypes held by the interviewer and therefore used naive subjects, that is, college students, with no interviewing experience.

A methodological shortcoming with this type of study, pointed out by Hollmann (1972), is that the information units are neither independent nor show equivalent, symmetrical patterns of intercorrelation. To counter this, he asked raters for a base rate of success for the company and a probability of success after each item of information was presented. This procedure might be altered to distinguish the ideal applicant and expected typical applicant by asking interviewers for base rates of success of each stereotype applicant and analyzing error scores of each bit of information compared to each base rate. This is a matter for future experimentation.

Finally, there is the problem that the contrast effects for unfavorable information account for less than 1% of the variance. Wexley et al. (1972) point out that this problem may be due to possible ceiling and floor effects caused by the extreme levels of suitability or that these extreme levels do not represent ambiguous stimuli and hence there was no opportunity for selective perception. Modifications in the stimulus materials used in this study might lead to shifts in the percentage of accounted variance.

In summary, this study has delineated the effects of ideal applicants, expected typical applicants, and favorability and orderings of information on rater's impressions. It has reviewed the importance of yet another source of variance, the expected typical applicant, in interviewer's decisions.

REFERENCES

BRISCOE, M. E., WOODYARD, H. D., & SHAW, M. E. Personality impression change as a function of the favorableness of first impressions. *Journal of Personality,* 1967, **35,** 343–357.

CARLSON, R. E. Effect of interview information in altering valid impressions. *Journal of Applied Psychology,* 1971, **55,** 66–72.

HAKEL, M. D. Similarity of post-interview trait rating intercorrelations as a contributor to interrater agreement in a structured employment interview. *Journal of Applied Psychology,* 1971, **55,** 443–448.

HAKEL, M. D., HOLLMANN, T. D., & DUNNETTE, M. D. Accuracy of interviewers, certified public accountants, and students in identifying the interests of accountants. *Journal of Applied Psychology,* 1970, **54,** 115–119.

HAKEL, M. D., OHNESORGE, J. P., & DUNNETTE, M. D. Interviewer evaluations of job applicants' resumés as a function of the qualifications of the immediately preceding applicants: An examination of contrast effects. *Journal of Applied Psychology,* 1970, **54,** 27–30.

HASTORF, A. H., SCHNEIDER, D. J., & POLEFKA, J. *Person perception.* Reading, Mass.: Addison-Wesley, 1970.

HAYS, W. L. *Statistics.* New York: Holt, Rinehart and Winston, 1963.

HOLLMANN, T. D. Employment interviewers' errors in processing positive and negative information. *Journal of Applied Psychology,* 1972, **56,** 130–134.

LONDON, M. The effects of the interviewer's ideal applicant stereotype and expected typical applicant and type, order, and favorability of information on impression formation in the selection interview. Unpublished master's thesis, the Ohio State University, 1972.

MAYFIELD, E. C., & CARLSON, R. E. Selection interview decisions: First results from a long-term research project. *Personnel Psychology,* 1966, **19,** 41–53.

ROWE, P. M. Individual differences in selection decisions. *Journal of Applied Psychology,* 1963, **47,** 304–307.

WEBSTER, E. C. *Decision making in the employment interview.* Montreal: Eagle, 1964.

WEXLEY, K. N., YUKL, G. A., KOVACS, S. Z., & SANDERS, R. E. Importance of contrast effects in employment interviews. *Journal of Applied Psychology,* 1972, **56,** 45–48.

(Received November 8, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 2, 163–167

COMPARABILITY OF INTERVIEWER, GROUP, AND INDIVIDUAL INTERVIEW RATINGS

LARRY F. MOORE,[1] AND ALEC J. LEE

Faculty of Commerce and Business Administration
University of British Columbia

Interview trait ratings made by individual and groups of managers following exposure to videotaped interview playbacks were compared to ratings made by experienced interviewers participating in 34 live interviews couched in a typical entry-level managerial employment setting. A convergent and discriminant validity analysis was made to assess the adequacy of the interview rating procedure. Mean ratings made by groups of managers were similar to mean ratings made by the interviewers. Agreement between the two dissimilar rater types using direct (face to face) and indirect (videotaped) stimuli indicated that perceptual distortion was not a strong factor. Individual managers' mean ratings suggested a leniency effect.

Individual differences among interviewers have been shown to account for some of the overall differences in interview judgments and sources of interviewer–decision bias that have been documented (Ulrich & Trumbo, 1965; Wright, 1969). A common strategy for countering the effect of individual differences in decision making is to combine individuals into groups. Group versus individual decision making and problem solving has been studied in a wide variety of settings and has been extensively reviewed (Lorge, Fox, Davitz, & Brenner, 1958; Maier, 1967; Hall & Williams, 1970). With some notable exceptions (Bouchard & Hare, 1970; Dunnette, Campbell, & Jaastad, 1963; Campbell, 1968) most of the research supports group over individual judgment.

Although interviewing by groups (panel interviewing) has been used in military and industrial settings (Office of Strategic Services, 1948; Taft, 1959) the effectiveness of group compared to individual decision making has not been examined in the context of the employment interview. One feasible method for achieving group decision making in the employment interview is to record the interview on videotape for later playback to a group of judges who make a collective decision. The purpose of the present study was to compare ratings made by individuals and groups of managers following exposure to videotape recorded interview playbacks with ratings made by experienced interviewers participating in live interviews in a typical entry-level managerial employment setting.

METHOD

Research Site and Subjects

The interviews forming the basis of the study were conducted at the western headquarters of a large Canadian commercial bank. Thirty-four graduating students of a university business school served as interview subjects. The median age of the subjects (all male) was 23 years and their mean grade level was B with a range from C to A−. The mean number of previous employment interviews experienced by the subjects was 5.5 and all had undergone at least one interview with another company.

Each subject participated in a 25–30 minute semi-structured employment interview (initial screening) with an interviewer from the bank's staff. Six interviewers each interviewed five or six different subjects. All interviewers had been conducting college employment interviews as a major part of their work roles during at least the previous six months and all had received either a one- or a two-week formal course on interview methods. Thus, the interviewers were considered to be experienced and trained in a manner typical of many business firms. The interviewers followed the bank's usual interview format designed to elicit attitudinal, behavioral, and factual information not attainable elsewhere in the selection process and to provide the interviewees with information designed to enhance the bank's image. All 34 interviews were recorded on one-half inch videotape using split screen staging to allow a head and upper body playback presentation of the interviewer and interviewee simultaneously.

---

[1] Requests for reprints should be sent to Larry F. Moore, Faculty of Commerce and Business Administration, University of British Columbia, Vancouver 8, Canada.

TABLE 1

MEANS, STANDARD DEVIATIONS, AND SIGNIFICANT DIFFERENCES FOR INTERVIEWERS', GROUPS', AND INDIVIDUALS' RATINGS

Trait	Ratings								
	Interviewers			Groups			Individuals		
	$\bar{X}$	*SD*	I vs. G	$\bar{X}$	*SD*	G vs. I′	$\bar{X}$	*SD*	I′ vs. I
Attitude	3.20	.69		3.35	.65		3.53	.61	2.05*
Appearance	2.94	.49		3.06	.42		3.27	.45	2.85**
Interest	2.85	.61		3.00	.78		3.12	.59	
Intelligence	3.12	.33	2.87**	3.40	.50		3.47	.51	3.41**
Leadership	3.12	.54		3.15	.70		3.27	.57	
Maturity	3.20	.48		3.27	.62	2.29*	3.62	.65	2.97**
Motivation	3.15	.56		3.12	.81		3.27	.57	
Persuasiveness	3.00	.43		3.15	.66		3.15	.66	
Self-confidence	3.17	.52	2.02*	3.44	.56		3.41	.61	
Self-expression	3.17	.58		3.21	.59		3.50	.66	2.15*
Sociability	3.17	.52		3.32	.53		3.44	.66	
Potential	2.97	.58		2.85	.86		3.09	.71	
Overall	3.00	.43		3.12	.69		3.27	.57	2.18*

Note. *t* test (two-tailed) is for significant difference. Abbreviations: I = interviewers, G = groups, and I′ = individuals.
* $p < .05$.
** $p < .01$.

The subjects were rated on 5-point graphic scales with equally spaced intervals labeled excellent, superior, average, marginal, and unsatisfactory. Judgments were made on 12 dimensions (traits): attitude, appearance, interest, intelligence, leadership, maturity, motivation, persuasiveness, self-confidence, self-expression, sociability, and potential. Each trait to be rated was amplified by definitional phrases or sentences. A separate overall rating was made using the 5-point graphic scale.

For each interview, three ratings were completed on each of the 13 traits. The first rating was made by each interviewer immediately following the live interview. The second rating was made by an individual manager in the bank following videotape playback of the interview. The third rating consisted of the pooled judgments of three managers based on group discussion following videotaped playback of the interview. Twenty-four line managers agreed to participate in the videotape rating exercise. Six groups of three persons and six individuals each viewed and rated five or six different interviews. All managers were employees of the bank with a mean of 18 years experience and all had intermittent experience in interviewing and selecting candidates for employment. Viewers participating as individual raters were not used for any of the group ratings and no rater viewed the same interviewer or interview on tape more than once. No restrictions were made on the rating groups in postreplay discussion. Prior to and during the interview, interviewers and other raters had access to and used the candidate's completed application form in forming their judgments. Thus, in realistic fashion, the study examined ratings based on interviews plus ancillary data. An investigation of the completed standardized application blanks did not reveal any systematic differences in interviewee backgrounds likely to bias the rater comparisons. To test for intrasample bias, one-way analyses of variance were calculated for the ratings made by each interviewer, group and individual in each of the five or six successive interviews rated (overall trait rating only). No significant intrasample differences were found (interviewers $F = .74$; groups $F = 1.67$; individuals $F = .66$, $df = 5/28$).

Mean and standard deviations were computed for the 12 interview traits and for the overall trait scores derived from each rating method, then the means were compared. A convergent and discriminant validity analysis (Campbell & Fiske, 1959) was made to compare more fully the three interview rating methods and to examine the adequacy of the general interview technique used by the bank hosting the study.

RESULTS

Table 1 presents means and standard deviations of the interviewer, group, and individual ratings of the interview subjects. Mean trait ratings by interviewers following live interviews agreed closely with the mean ratings of groups following videotape playback and discussion. The mean interview ratings

TABLE 2

CONVERGENT AND DISCRIMINANT VALIDITIES OF INTERVIEWERS-GROUPS, INTERVIEWERS-INDIVIDUALS, AND INDIVIDUALS-GROUPS

Trait	Heteromethod block								
	Interviewers-groups			Interviewers-individuals			Individuals-groups		
	Validity coefficient	Highest heterotrait value	No. heterotrait values higher[a]	Validity coefficient	Highest heterotrait value	No. heterotrait values higher[a]	Validity coefficient	Highest heterotrait value	No. heterotrait values higher[a]
Attitude	31*	56	7	38*	56	6	28*	59	7
Appearance	46*	39	0	21	41	9	24	39	3
Interest	51*	47	0	39*	41	1	13	48	16
Intelligence	25	48	12	02	34	20	05	42	20
Leadership	60*	40	0	09	54	12	28*	52	2
Maturity	32*	44	9	36*	57	4	11	43	17
Motivation	16	53	15	26	57	11	26	48	9
Persuasiveness	32*	53	11	00	40	19	30*	42	3
Self-confidence	45*	41	0	15	32	14	34*	42	7
Self-expression	51*	60	1	24	45	6	04	46	22
Sociability	44*	60	7	39*	45	2	01	64	24
Potential	42*	42	0	30*	57	10	37*	64	12
Overall	10	53	21	38*	56	5	31*	43	10
Median	42	48		26	45		26	46	

[a] Number of heteromethod off-diagonal coefficients in corresponding row and column higher than validity coefficient; maximum = 24.

* $p < .05$.

were significantly higher than the group ratings for intelligence and self-confidence. Individual managers differed from the interviewers to a greater extent. The mean individual ratings were significantly higher than the mean interviewer ratings for six traits. Inspection of the mean interviewer ratings show them to be uniformly lower than the mean individual ratings and lower than all of the corresponding mean group ratings except motivation and potential. This pattern may reflect a tendency reported by Rowe (1963) for experienced interviewers to be more critical.

Table 2 summarizes the analysis of convergent and discriminant validity following the Campbell and Fiske (1959) procedure. Essentially, the procedure enables a detailed analysis of intercorrelations of trait scores among and between traits within and among methods when more than one method is used to measure more than one trait. The Campbell–Fiske design provides one indication of convergent validity and three indications of discriminant validity.

Convergent validity is indicated when two or more independent measures tend to agree in the measurement of a given variable (trait). In Table 2 the correlations between methods are shown in columns 1, 4, and 7. Minimal requirements for convergent validity were met for 10 of 13 traits in the interviewer-group comparison (r significantly different from 0); however, the median validity coefficient was low. In the interviewer–individual and individual–group comparisons less than half the traits satisfied the criterion for convergent validity.

The first test for discriminant validity requires that the validity coefficient for a trait measure by two different methods should be higher than the correlations between that trait and other traits measured by different methods (i.e., having neither method nor trait in common). This criterion is met by five traits in the interviewer–group comparison but is not met by any trait in the interviewer–individual or individual–group comparisons.

The second test for discriminant validity requires that measures of a given trait made with independent methods correlate higher than correlations between the given trait and other traits when measured by a common method. This test indicates the strength of a trait across methods as well as its independence from other traits. All heteromethod trait intercorrelations were exceeded by at least 50% of the intertrait correlations within methods except for appearance, interest, and leadership, all within the interviewer–group heteromethod block. These were exceeded by 8%, 25%, and 11% respectively. Clearly, the second test for discriminant validity was not met. Intercorrelations between traits within the three monomethod blocks were quite high (for interviewers, median $r = .40$; for groups, median $r = .53$; for individuals, median $r = .45$), indicating that most of the traits themselves may have lacked definitional clarity.

An examination of the patterns of trait intercorrelations within and between rating methods for similarity constitutes the third test for discriminant validity. No pattern similarity could be identified, thus placing further doubt on the definitional clarity of the traits used by the bank.

DISCUSSION

The results indicated that mean trait ratings made by groups of managers after exposure to videotape playbacks of candidates' interviews are highly similar to those of interviewers having the typical level of training and experience. Perception biases often associated with interviewer judgment such as stereotyping, early closure, or data seeking operating singly or cumulatively did not appear to produce a differentiating effect. Hakel, Ohnesorge, and Dunnette (1970) have found that the effect of perception bias may be negligible in some instances.

The tendency for individual managers to rate relatively high suggests a leniency effect. Untrained raters have been shown to be less critical than experienced interviewers (Rowe, 1963). Although the individual managers used as raters in this study had been involved in interview decisions in the past, none had previously received formalized training in interviewing or decision making. Further research is needed to determine the precise way in which group discussion influences decision making in interview rating. Holloman and Hendricks (1971) have demonstrated the error-correcting propensity of social interaction in group decision making in other settings. Group discussion may influence untrained raters to be more critical in evaluating interviewees.

Interview reliability and validity analysis is facilitated via videotape because of the replay feature. Although design limitations precluded investigation of intrarater reliability in this study, tapes of three of the bank's screening interviews with associated application blanks were shown to groups of three ($n = 27$) and individual ($n = 104$) student judges at two viewing sessions separated by a one-week interval. Intrarater consistency differed considerably across the 13 interview variables but correlation coefficients as high as .70 to .80 were reached although the medians were .54 and .59 for groups and individuals respectively.[2]

The convergent and discriminant validity analysis raised serious questions about the interview process itself. The convergent validity coefficients were quite low throughout, indicating large amounts of unexplained variance between rating methods on most traits. The analysis for discriminant validity revealed generally high intertrait correlations within methods, but no common pattern existed among methods. Suspicion was thus cast on the traits themselves; their definitional clarity, degree of overlap or commonality, and their realism and relevance to the raters. Other researchers (Howell & Vincent, 1970; Hakel & Schuh, 1971) have begun to define interviewee attributes that are suitable for interview assessment. Further research is needed to determine the relative effects of all sources of undesirable influence on the interview and on the interview process. Some are apparently more critical than others.

---

[2] Further details of this reliability investigation will be supplied by the authors on request.

REFERENCES

BOUCHARD, T. J., & HARE, M. Size, performance, and potential in brainstorming groups. *Journal of Applied Psychology,* 1970, **54**, 51–55.

CAMPBELL, D. T., & FISKE, D. W. Convergent and discriminant validation by the multitrait multimethod matrix. *Psychological Bulletin,* 1959, **56**, 81–105.

CAMPBELL, J. Individual versus group problem solving in an industrial sample. *Journal of Applied Psychology,* 1968, **52**, 205–210.

DUNNETTE, M. D., CAMPBELL, J., & JAASTAD, K. The effect of group participation on brainstorming effectiveness for two industrial samples. *Journal of Applied Psychology,* 1963, **47**, 30–37.

Hakel, [illegible] D., Ohnesorge, J. P., & Dunnette, M. D. Interviewer evaluation of job applicants' resumes as a function of the qualifications of the immediately preceding applicants: An examination of contrast effects. *Journal of Applied Psychology*, 1970, **54**, 27–30.

Hakel, M. D., & Schuh, A. J. Job applicant attributes judged important across seven diverse occupations. *Personnel Psychology*, 1971, **24**, 45–52.

Hall, [illegible] & Williams, M. S. Group dynamics training and improved decision making. *Journal of Applied Behavioral Science*, 1970, **6**, 39–68.

Holloman, C. R., & Hendricks, H. W. Problem solving in different sized groups. *Personnel Psychology*, 1971, **24**, 489–500.

How[illegible], M. A., & Vincent, J. W. Factor analysis of interview data. *Journal of Applied Psychology*, 1970, **54**, 313–315.

Lorge, I., Fox, D., Davitz, J., & Brenner, M. A survey of studies contrasting the quality of group performance and individual performance, 1920–1957. *Psychological Bulletin*, 1958, **55**, 337–372.

Maier, N. R. F. Assets and liabilities in group problem solving: the need for an integrator's function. *Psychological Review*, 1967, **74**, 239–249.

Office of Strategic Services. *Assessment of men*. New York: Rinehart, 1948.

Rowe, P. M. Individual differences in selection decisions. *Journal of Applied Psychology*, 1963, **47**, 304–307.

Taft, R. Multiple methods of personality assessment. *Psychological Bulletin*, 1959, **56**, 333–352.

Ulrich, L., & Trumbo, D. The selection interview since 1949. *Psychological Bulletin*, 1965, **63**, 100–116.

Wright, O. R. Summary of research on the selection interview since 1964. *Personnel Psychology*, 1969, **22**, 391–413.

(Received October 10, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 2, 168–171

PARTICIPATION EFFECTS IN APPRAISAL INTERVIEWS CONDUCTED IN A TRAINING SITUATION[1]

JOSEPH M. HILLERY AND KENNETH N. WEXLEY[2]

University of Akron

Participation, order, and interviewee personality effects were studied in appraisal interviews conducted as a part of a training program. Results showed that nonparticipative appraisal interviews brought about greater behavior change and interviewee satisfaction than participative interviews. Order of presentation did not effect interview outcome, and personality variables were not found to moderate the effectiveness of type of interview. The importance of the training situation was discussed in relation to the results of the study.

The desirability of allowing subordinates to participate in their own performance appraisal interviews has been voiced for a number of years (e.g., Kindall & Gatza, 1963; Likert, 1959; McGregor, 1957). McGregor (1957) was the first to outline an entire performance appraisal approach in which the subordinate takes an active part. Studies which have experimentally manipulated the amount of participation in appraisal interviews (French, Kay, & Meyer, 1966; Wexley, Singh, & Yukl, 1973) have supported the notion of the beneficial aspects of participation. French et al. (1966) found that higher levels of participation produced improvements in the supervisor–subordinate relations. Wexley et al. (1973) showed that participative appraisal interviews resulted both in greater subordinates satisfaction with the interview and greater intention to improve future job performance.

While Maier (1973) has generally supported the participative type of appraisal interview, he believes there are situations in which a nonparticipative approach is more appropriate. Maier (1958) contended that the tell and sell method (a nonparticipative approach) has its greatest potential with new and inexperienced workers. Furthermore, Maier (1973) held that criticism, typically given in the nonparticipative but not in the participative interview, is most likely to be accepted by the subordinate when he is still learning the job. Acceptance of criticism is important since most workers react defensively to criticism, resulting in less effort to improve performance (Kay, Meyer, & French, 1965). The major purpose of this study was to see if the beneficial aspects of participation in appraisal interviews found with experienced workers also holds for appraisal interviews with trainees. Interview effectiveness was measured by performance improvement and interviewee satisfaction with the interview.

Two additional variables were examined in the appraisal process. First, personality factors of self-esteem, authoritarianism, and need for independence have been found to moderate the effectiveness of various styles of leadership (Kay et al., 1965; Vroom, 1959). However, not all studies investigating these variables have obtained similar results (Tosi, 1970; Wexley et al., 1973). For example, Wexley et al. (1973) found that it was desirable to allow a subordinate to participate in the appraisal process regardless of his personality structure. In the present study, first, all three personality variables and their relationship with interview effectiveness were examined. Second, order effects in the nonparticipative interviews were examined to see if it made any difference if the ineffective items were presented first, last, or sandwiched in between effective items.

METHOD

Procedure

Appraisal interviews were conducted as part of a teacher-training program known as microteaching.

---

[1] This study is based on a dissertation submitted by the first author to the Department of Psychology, University of Akron, in partial fulfillment of the requirement for the PhD degree.

The authors wish to thank Isobel Pfeiffer for her cooperation and assistance.

[2] Requests for reprints should be sent to Kenneth N. Wexley, Department of Psychology, University of Akron, Akron, Ohio 44325.

Microteaching is similar to an actual classroom situation except that it is scaled-down in terms of teaching time and size of class. It is a widely used (Allen & Ryan, 1969) and effective (Kallenbach & Gall, 1969) method of teacher training.

Four different ways of presenting feedback in the appraisal interview comprised the four experimental conditions. The four conditions included one participative and three nonparticipative types of appraisal interviews. The three nonparticipative interviews differed in the order in which the two effective and two ineffective behaviors were presented: (*a*) the effective behaviors presented first, then the ineffective (EI); (*b*) the ineffective then effective (IE); and (*c*) one effective, two ineffective, then the remaining effective (EIE or commonly called the "sandwich technique").

Each of 16 teachers was exposed to all four experimental feedback conditions presented in counterbalanced order. This within subjects design had a total of 64 teach–interview–reteach sequences (i.e., 4 sequences for each of the 16 teachers). The initial teaching session (teach) consisted of the teacher teaching three high school students a 10-minute lesson aimed at the completion of one task. After the teaching session, two supervisors were given 10 minutes to confer and decide which two effective and two ineffective behaviors were to be fed back to the teacher. The appraisal interview consisted of a 10-minute discussion between one supervisor and the teacher using one of the four types of interviews. The teacher was then given a half hour to prepare for his reteach sequence. The reteach sequence consisted of the teacher reteaching the same 10-minute lesson, but to a different group of students.

After every teach and reteach sequence, two supervisors filled out the behavior checklist for that task. The checklists were scored using the formula $E - I/E + I$, where the Es are the effective behaviors checked and the Is are the ineffective behaviors checked. After each appraisal interview, the teacher filled out a graphic rating scale indicating how satisfied he was with the interview.

Subjects

Three groups of subjects were used: (*a*) 16 undergraduate education majors who taught the lessons (teachers); (*b*) 25 graduate students who observed the performance of the teachers and who conducted the interviews (supervisors); (*c*) 12 high school students who were taught the lessons (students).

The teachers were 16 undergraduate education majors enrolled in an undergraduate course. The supervisors were experienced elementary and secondary teachers enrolled in a graduate course in supervision. The microteaching experience was included in the training of the teachers and supervisors. The students being taught the lessons were 12 high school students paid to participate in the project.

Materials

Self-esteem. The instrument used consisted of a set of 13 bipolar adjectives. Each teacher twice filled out a 13-bipolar-adjective questionnaire, once for how he perceived himself as a teacher and once for how he perceived the ideal teacher. A person's self-esteem score was computed by determining the discrepancy between his actual and ideal scores.

Authoritarianism. The degree of teacher authoritarianism was measured by responses to a 25-item scale developed by Adorno, Frenkel-Brunswick, Levinson, and Sanford (1950) and subsequently adapted and used by Vroom (1959). Adorno et al. report reliability studies of this instrument with average reliability coefficients of .90.

Need for independence. The teachers' need for independence was measured by responses to a 16-item questionnaire used by Vroom (1959) and adapted from a larger instrument employed by Tannenbaum and Allport (1956). Vroom reported that the test-retest reliability over a seven-month period of a shortened form of the questionnaire was .61 for 90 supervisors.

Checklists. Five separate checklists were developed, one for each of the five teaching tasks. Each checklist contained 20 behavior specimens, 10 effective and 10 ineffective. The items were obtained from 20 experienced teachers familiar with each of the five teaching tasks. Additional behavioral items were adapted from the evaluation guides published with the teaching tasks (Davis, Gregory, Kyselka, Morse, & Smoot, 1970). To estimate the reliability of the checklists, a ratio was computed between the number of instances in which two raters agreed (both raters checked or failed to check an item) to the total number of agreements plus disagreements. The percentage of agreement for the overall checklist was 80%, while the percentage for the checklists scored only for the four fed-back items was 76%.

Teaching tasks. The five tasks, taken from Davis et al. (1970), included clarity of instructional objectives, presentation, interaction, questioning, and specific refocusing. These tasks were considered to be important dimensions of effective teacher performance.

Satisfaction. The teacher's satisfaction with the appraisal interview was measured by averaging his responses to the following three questions: (*a*) How effective was the conference in covering the important aspects of your "teach" behavior? (*b*) If you were to be the supervisor in an ideal conference, how similar would it be to the conference you just had? (*c*) How satisfied were you with the conference?

RESULTS

Table 1 presents the means and standard deviations for the participative and nonparticipative interviews on each of the following three dependent variables: (*a*) the checklist scored for all behaviors checked (overall

TABLE 1

MEANS AND STANDARD DEVIATIONS FOR PARTICIPATIVE AND NONPARTICIPATIVE INTERVIEWS ON THREE DEPENDENT VARIABLES

Dependent variable	$\bar{X}$	SD
Overall checklist		
participative	.04	.27
nonparticipative	.29	.10
Fed-back checklist		
participative	.17	.42
nonparticipative	.48	.17
Satisfaction		
participative	3.60	.72
nonparticipative	4.21	.43

checklist), (*b*) the checklist scored only for those behaviors which were fed back to the teacher (fed-back checklist), and (*c*) the overall measure of the interviewee's satisfaction with his interview. The scores for the overall checklist and fed-back checklist represent gain scores from pretest (teach) to posttest (reteach). The nonparticipative interview scores are the gain scores for all teachers averaged across the three nonparticipative interviews. Analyses of variance revealed that the nonparticipative interviews brought about significantly greater performance increases on the overall checklist ($F = 9.53$, $df = 1/15$, $p < .01$) and the fed-back checklist ($F = 7.59$, $df = 1/15$, $p < .05$) than did the participative interviews. Also, the interviewees were significantly more satisfied ($F = 12.72$, $df = 1/15$, $p < .01$) with the nonparticipative interviews than with the participative ones.

Results of the comparisons examining order effects indicated that no one order of presentation was found more effective or satisfying than any other presentation order. This finding is contrary to the advice often given in business journals that one order of presentation is superior to another (e.g., Leskovec, 1967).

Differences in teacher self-esteem and authoritarianism showed no main effects and no interactions with the amount of participation or the order of presentation. On one dependent variable (the overall checklist) there was a main effect for teachers' need for independence ($p < .05$). Those persons who had a low need for independence showed significantly more improvement than did those persons who had a high need for independence. In no case, however, did the need for independence interact with the amount of participation or the order of presentation.

DISCUSSION

From the results, it is clear that the nonparticipative interviews were more effective than participative interviews in producing improved teaching as well as resulting in higher teacher satisfaction with the interview. These results do not support the previous experimental studies examining appraisal interviews (French et al., 1966; Wexley et al., 1973) that have consistently shown the superiority of participative approaches. The most notable difference between these two studies and the present one is that the two previous studies dealt with real-life appraisals conducted with experienced workers (French et al., 1966) or with students role playing experienced workers (Wexley et al., 1973), while the present study involved inexperienced people in a training setting. The teachers in the present study were desirous of being evaluated by the supervisor and were noticeably disappointed when asked to participate in their own evaluation. On several occasions, the teachers asked the supervisors for direct feedback after a participative interview had been conducted. This was indicative of the teachers' dissatisfaction and frustration with the participative approach in the training situation. The results from this study lend empirical support to Maier's (1958) notion that a tell and sell appraisal interview (i.e., nonparticipative) has its greatest effect with young and inexperienced workers.

The personality variables of self-esteem, authoritarianism and need for independence were not found to moderate the effectiveness of the participative and nonparticipative approaches. These results conflict with the findings of Vroom (1959) and Kay et al. (1965), but support the recent findings of Tosi (1970) and Wexley et al. (1973). The difference in results might be explained by either the different instruments used to measure the personality variables or the totally different situation in which the studies were conducted. Clearly, more research is needed to delineate the conditions under which personality factors are important in determining appraisal interview outcomes.

One limitation of the present study should be noted. The performance improvement was measured over a one-hour time period. Studies have demonstrated that the beneficial effects of training may not persist over a long period of time (Fleishman, 1953; Ivancevich, 1972). Additional research studying the appraisal interview in training using follow-up measures is clearly desirable. The microteaching setting, however, did afford a well-controlled situation in which actual behavior changes could be evaluated in a pretest-posttest design.

REFERENCES

Adorno, T., Frenkel-Brunswick, E., Levinson, D. J., & Sanford, N. *The authoritarian personality*. New York: Harper, 1950.

Allen, D., & Ryan, K. *Microteaching*. Reading, Mass.: Addison-Wesley, 1969.

Davis, O. L., Jr., Gregory, T. B., Kyselka, M. L., Morse, K. R., & Smoot, B. R. *Basic teaching tasks: A teaching laboratory manual for beginning teacher candidates*. Research and Development Center for Teacher Education, University of Texas at Austin, 1970.

Fleishman, E. A. Leadership climate, human relations training, and supervisory behavior. *Personnel Psychology*, 1953, **6**, 205–222.

French, J. R. P., Jr., Kay, E., & Meyer, H. H. Participation and the appraisal system. *Human Relations*, 1966, **19**, 3–20.

Ivancevich, J. M. A longitudinal assessment of management by objectives. *Administrative Science Quarterly*, 1972, **17**, 126–138.

Kallenbach, W. W., & Gall, M. D. The comparative effects of microteaching and a conventional training program on the classroom performance of elementary intern teachers. *Journal of Educational Research*, 1969, **63**, 136–141.

Kay, E., Meyer, H. H., & French, J. R. P., Jr. The effect of threat in a performance appraisal interview. *Journal of Applied Psychology*, 1965, **49**, 311–317.

Kindall, A. F., & Gatza, J. Positive program for performance appraisal. *Harvard Business Review*, 1963, **41**, 153–166.

Leskovec, E. A guide for discussing the performance appraisal. *Personnel Journal*, 1967, **46**, 150–152.

Likert, R. Motivational approach to management development. *Harvard Business Review*, 1959, **37**, 75–82.

Maier, N. R. F. *The appraisal interview: Objectives, methods, and skills*. New York: Wiley, 1958.

Maier, N. R. F. *Psychology in industrial organizations*. Boston: Houghton Mifflin, 1973.

McGregor, D. An uneasy look at performance appraisal. *Harvard Business Review*, 1957, **35**, 89–94.

Tannenbaum, A., & Allport, F. H. Personality structure and group structure: An interpretative study of their relationship through an event-structure hypothesis. *Journal of Abnormal and Social Psychology*, 1956, **53**, 272–280.

Tosi, H. A reexamination of personality as a determinant of the effects of participation. *Personnel Psychology*, 1970, **23**, 91–99.

Vroom, V. Some personality determinants of the effects of participation. *Journal of Abnormal and Social Psychology*, 1959, **59**, 322–327.

Wexley, K. N., Singh, J. P., & Yukl, G. A. Subordinate personality as a moderator of the effects of participation in three types of appraisal interviews. *Journal of Applied Psychology*, 1973, **58**, 54–59.

(Received February 12, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 2, 172–178

EXTENSIONS OF A PATH-GOAL THEORY OF MOTIVATION[1]

MARTIN G. EVANS[2]

Faculty of Management Studies, University of Toronto, Canada

This article extends and replicates the Evans's hypothesis concerning the way in which the behavior of the superior affects the subordinate's perceptions of expectancies and instrumentalities in the path-goal theory of motivation. Of the three moderators (the subordinate's locus of control, the subordinate's position in a web of role relationships, and the supervisor's upward influence) hypothesized, only the first was found to moderate the superior/subordinate relationship as predicted. Results for the other two moderators were equivocal. Additional implications of the path-goal model were explored, that is, the role of motivation as: (*a*) an intervening variable between supervisory behavior and subordinate behavior and (*b*) as a moderator in the behavior/satisfaction relationship.

The seminal article of Georgopoulos, Mahoney and Jones (1957) extended the expectancy theories of motivation (Lewin, 1938; Tolman, 1932) to the organizational scene. The model proposed was formalized and further extended by Vroom (1964) who made a more extensive analysis of the path-goal issue. He introduced two basic probability concepts: (*a*) Instrumentality, which provides a probability link between two outcomes, represents an individual's beliefs that one outcome is associated with another, for example, that being a high performer results in higher pay. (*b*) Expectancy, which provides a probability link between behavior and an outcome, represents an individual's beliefs that an outcome is associated with his behavior, for example, that increased effort results in higher performance. Vroom suggested that motivation to perform is made up of two components combined multiplicatively: (*a*) the valency of high performance, the attractiveness of the outcome (which in turn is comprised of the sum, over goals, of the product of the importance of each goal and the instrumentality of high performance for the goal) and (*b*) the path-goal expectancies that effort or working hard leads to high performance. Furthermore, following Georgopoulos et al. (1957), work behavior is a function of this motivational force, subject to a number of constraints such as the individual's ability and the freedom he has to vary his behavior.[3]

Variants and extensions of Vroom's model have been tested and confirmed in a variety of studies—experimentally (Graen, 1969); cross-sectionally (Gavin, 1970; Porter & Lawler, 1968; Schuster, Clark, & Rogers, 1971); longitudinally (Lawler, 1968). In addition, Porter and Lawler (1968), Graen (1969) and Evans (1970b) have further extended the model to show the relationship with job satisfaction. The path-goal theory provides a rather complete framework for understanding motivation in organizational settings.

Given that the path-goal framework has some support as a model of individual motivation in organizations, how and when do organizational factors affect motivation? Recent literature has suggested that organizational factors may affect motivation through the impact that these factors have on instrumentalities and expectancies. Compensation, promotion, and participation (Galbraith, 1968), supervisory behavior (Evans, 1970b), task complexity (House, 1971), and work group behavior (Graen, 1969) have all been reported to have an effect on these probabilities. In addition, interactions of some of these factors with one another and with the individual's own characteristics may also influence motivation. For example, House (1971) suggested complex interactions between supervisory behavior, task structure and the individual's tolerance for ambiguity.

---

[1] The research reported in this article was supported by a Senior Faculty Research award from the Seagram Foundation, administered by the Canadian Association of Schools of Business.

The author wishes to thank D. T. Hall, R. Mansfield, T. R. Mitchell and R. Payne for their comments on an earlier draft of this article.

[2] Requests for reprints should be sent to Martin G. Evans (formerly at the London Graduate School of Business, England), who is now at Faculty of Management Studies, University of Toronto, Toronto M5S 1V4, Ontario, Canada.

[3] While Georgopoulos et al. (1957) found their hypothesized relationships were enhanced when the individual had freedom from barriers to vary production, research is needed to explore whether or not these barriers are implicitly taken into account by the individual when he reports his expectancies or instrumentalities.

Evans (1970a, 1970b) has used the path-goal approach to link leader behavior with employee motivation and job satisfaction, and has proposed that two conditions were necessary for the supervisor to create a situation in which subordinates perceived strong links between high-performance paths and their idiosyncratic goals: (*a*) that the supervisor offered a range of valued rewards to the subordinate (thus showing consideration), and (*b*) that he made certain that the distribution of the reward was made contingent on performance (initiating structure). Evans's linkage model (following the path-goal theory) further suggests that when the goals were important to the subordinate and he saw a strong link between performance and his goals then he tends to be a high performer. When the individual was a high performer, and there was an actual link between performance and reward, then the individual attains his goals and experiences job satisfaction. The model implies that the independent variables of the leader behaviors only relate to the dependent variables, performance and satisfaction, if (*a*) the independent variable (leader behaviors) relates to a set of intervening variables (path-goal links) and (*b*) the intervening variables (path-goal links) relate to the dependent variables (performance and satisfaction). Research in two organizations tested this model (Evans, 1970b) and provided some support.

However, there were differences between the two organizations studied, and Evans (1970a) suggested three factors might moderate the supervisory behavior/subordinate motivation relationship: the upward influence in the hierarchy exhibited by the supervisor (Pelz, 1952); the role conflict experienced by the subordinate; and the extent to which the subordinate was internal or external in locus of control orientation (Rotter, 1966), or the generalized expectancies of the subordinate that reward and punishment were based on his own behavior rather than on random occurrences.

The present study was designed to test these speculations as well as to replicate and extend the earlier research (Evans, 1970b). The following hypotheses were made concerning the relationship between supervisor behavior and motivation: *Hypothesis 1*—The more considerate the supervisor, the higher the motivation is for high performance paths. The greater the upward influence of the supervisor, the more likely that he could offer and deliver many rewards (consideration) to the employees, thus influencing their expectancies and increasing their motivation. *Hypothesis 2*—The greater the initiation of structure, the higher the motivation is for high performance paths. The higher the role conflict experienced by the subordinates (in that different role senders had clearly different expectations of how the individual should behave), the less effect the supervisor has on making the distribution of the rewards contingent on performance (initiating structure). The more internally controlled the subordinate, the stronger the relationship. *Hypothesis 3*—Motivation is an intervening variable in the leadership/behavior relationship. *Hypothesis 4*—The higher the motivation (Σ Valence × Expectancy) to undertake a particular activity, the more frequently the activity is undertaken. *Hypothesis 5*—Motivation is a moderator of the behavior/satisfaction relationship.

METHOD

Subjects

Subjects were 86 young managers enrolled in the first year of a part-time master's of business administration (MBA) program at the University of Toronto. Median age was 25–29, and most subjects were university graduates. All subjects had been with their present companies for about two years in a variety of functional specialities. Most subjects (57%) had been brought up in Southern Ontario,

TABLE 1

CORRELATIONS BETWEEN INITIATION OF STRUCTURE, CONSIDERATION, UPWARD INFLUENCE, ROLE CONFLICT, ROLE AMBIGUITY AND INTERNAL–EXTERNAL ORIENTATION

Item	1	2	3	4	5	6
1. Initiation	—					
2. Consideration	.54***	—				
3. Upward influence	.63***	.59***	—			
4. Role ambiguity[a]	.12	.51***	.17	—		
5. Role conflict[a]	.14	.48***	.29**	.67***	—	
6. Internal-external[b]	−.26**	−.24*	−.26**	−.24*	−.25*	—
$\bar{X}$	105.70	105.84	36.16	34.35	39.36	8.36
SD	14.88	18.39	6.30	5.69	7.98	3.98

[a] High scores imply lower conflict and ambiguity.
[b] Higher score implies a more external orientation.
* $p \leq .05$.
** $p \leq .01$.
*** $p \leq .001$.

52% of them in large urban centers (i.e., the Toronto/Hamilton, Ontario, conurbation).

Measuring Instruments

Leadership Behavior. The subordinate's perceptions of leadership behavior were measured with Version XII of the Leader Behavior Description Questionnaire (Stogdill, 1968). "Upward influence" was measured with the "superior orientation" subscale; "consideration" was measured by the sum of the consideration, tolerance for freedom, and integration subscales; initiation was measured by the sum of the structure, role assumption and production emphasis subscales.

Role conflict. This was measured with the paper and pencil questionnaire developed by Rizzo, House and Lirtzman (1970).

Powerlessness. This was measured with Rotter's (1966) Internal-External Control scale, which has had extensive experimental validation (Rotter, 1966; MacDonald, 1969).

Job performance. Self-rated measures of the following were obtained: effort, performance, frequency of making suggestions to the superior, frequency of following the superior's instructions, frequency of helping fellow workers, and frequency of training subordinates.

Path-goal expectancies and instrumentalities. These were measured by questions designed to tap: (*a*) expectancies whether or not effort leads to performance and (*b*) instrumentality whether or not high performance resulted in reward of some kind.

Goal attainment. This was measured with a version of Porter's (1961) instrument, tapped the facets of self-actualization, self-esteem, autonomy, social, security, pay, supervision, promotion, and family. The family facet was included to try to explore a non-work-related goal that might conflict strongly with the work-related goals. In addition, a managerial version of the Job Descriptive Index (JDI) (Warr & Routledge, 1969) was used to measure satisfaction with firm, job, pay, promotion opportunities, superior, peers, and subordinates.

Overall satisfaction. This was measured with the Brayfield-Rothe (1951) instrument.

One major limitation of this study lies in the fact that the respondents were the source of all data. It is possible that individuals may have been lenient in rating their behavior, which restricts the range of the dependent variables.

RESULTS

Relationship Between Initiation of Structure and Consideration with Motivational and Performance Variables

It was hypothesized (Hypotheses 1 and 2) that high consideration and high initiation is associated with high motivation. However, a problem emerged in testing these hypotheses. For this sample, perceived initiation and consideration were highly correlated ($r = .54$, $p < .001$); see Table 1; hence, the original intention of analyzing the data through a two-way analysis of variance (ANOVA) was inappropriate. Accordingly, the technique of multiple regression was employed—motivation and performance measures[4] were the dependent variables, while initiation, consideration, superior orientation, I–E control, role ambiguity, and role conflict were employed as predictors. The results are presented in Table 2.

---

[4] Additionally, the product of instrumentalities and expectancies (EPI)—a measure of motivation unweighted by the importance of goals—was included (for each activity) as a dependent variable.

TABLE 2

M LE REGRESSION OF MOTIVATION AND PERFORMANCE WITH CONSIDERATION (C), INITIATION OF STRUCTURE (I), UPWARD INFLUENCE (UI), INTERNAL-EXTERNAL CONTROL (I-E), ROLE CONFLICT (RC), ROLE AMBIGUITY (RA).

Dependent variable	*R*	% Variance accounted for by independent variables					
		I	C	I-E	RC	RA	UI
PI-high performance	.60		29.3	−6.4			
Motivation-high performance	.63		31.4	−8.0			
PI train subordinates	.60		27.6	−5.0	−4.0		
Motivation train subordinates	.64		29.6	−7.0	−4.6		
PI help follows	.62		27.1	−5.6		−5.3	
Motivation help follows	.65		28.6	−7.2		−6.0	
PI suggestions to boss	.63		28.3	−2.9	−3.4		5.7
Motivation suggestions to boss	.66		30.4	−3.4		−4.3	6.1
PI do what boss says	.57		24.5	−4.5			3.0
Motivation do what boss says	.54		4.0				25.2
Effort	.41		11.9	−5.2			
Do what boss says	.38	8.9			5.2		
Help fellows	*ns*						
Performance	*nd*						
Train subordinates	.43	11.9				6.8	
Suggestions to boss	.30		8.9				

First, consideration consistently appeared as being positively related to motivation[5] and to some of the performance variables; second, internals had higher motivation than externals; third, in some cases motivation was higher when role conflict or ambiguity was higher.[6]

In order to test Hypotheses 1 (second part), 2 (second and third parts), the regression was run after splitting the sample on the three control variables: superior's upward influence, internal-external control, and role conflict. The results for these are summarized below in Table 3.[7]

Upward influence as a moderator. The effect of this variable on the relationship between the dependent variables and consideration, with which upward influence is correlated ($r = .59$, $p < .001$), is inconsistent. The major effect of controlling for upward influence is that the relationship with consideration diminishes; what relations do remain are

TABLE 3

NUMBER OF SIGNIFICANT MULTIPLE CORRELATIONS BETWEEN MOTIVATION AND PERFORMANCE AND CONSIDERATION WHEN CONTROLLING FOR UPWARD INFLUENCE, INTERNAL-EXTERNAL CONTROL, AND ROLE CONFLICT

Variable	Variable	
	Motivation? EPI ($n = 10$)	Performance ($n = 6$)
No control	9	2
High-upward influence	4	1
Medium upward influence	2	1
Low-upward influence	3	1
High-role conflict	8	1
Medium role conflict	0	4
Low-role conflict	5	1
Internals	10	1
Externals	1	1

Note. EPI = product of instrumentalities and expectancies.

---

[5] It should be noted that the motivation measures are not independent from each other. All contain a component of the valence of high performance work.

[6] It should be noted that, contrary to the findings of Graen (1969) and Evans (1970a, 1970b), the correlations between the independent variables and motivation are slightly higher than the correlations between the independent variables and path-goal expectancies for four out of five behaviors.

[7] The tables on which these comments are based can be obtained from the author.

TABLE 4

CORRELATION BETWEEN CONSIDERATION, MOTIVATION AND PERFORMANCE, INCLUDING PARTIALLING MOTIVATION OUT OF THE CONSIDERATION-PERFORMANCE RELATIONSHIPS

Effort	1	2	3	4
1. Effort	—			
2. Consideration	.34**	—		
3. EPI high performance	.43***	.54***	—	
4. Motivation high performance	.44***	.56***	.98***	—
E with C/EPI	.15			
E.C/motivation	.13			
Do what boss says	**1**	**2**	**3**	**4**
1. Do what boss says	—			
2. Consideration	.26**			
3. EPI do what boss says	.24**	.49***		
4. Motivation do what boss says	.21*	.46***	.98***	
DBS.C/EPI	.22*			
DBS.C/motivation	.24**			
Help fellows	**1**	**3**	**2**	**4**
1. Help fellows	—			
2. Consideration	.20*	—		
3. EPI help fellows	.19*	.52***	—	
4. Motivation help fellows	.21	.54***	.98***	—
HF.C/EPI	.12			
HF.C/motivation	.11			
Train subordinates	**1**	**2**	**3**	**4**
1. Train subordinates	—			
2. Consideration	.31**	—		
3. EPI train subordinates	.33**	.53***	—	
4. Motivation train subordinates	.37***	.54***	.98***	—
TS.C/EPI	.18			
TS.C/motivation	.14			
Suggest to boss	**1**	**2**	**3**	**4**
1. Suggest to boss	—			
2. Consideration	.30**	—		
3. EPI suggest to boss	.14	.53***	—	
4. Motivation suggest to boss	.14	.55***	.98***	—
SB.C/EPI	.27*			
SB.C/motivation	.27*			

Note. EPI = product of instrumentalities and expectancies; E = effort; C = consideration; DBS = do what boss says; HF = help fellows; TS = train subordinates; SB = suggest to boss.

* $p < .05$.
** $p < .01$.
*** $p < .001$.

distributed over all three levels of upward influence. There is, therefore, little support for Hypothesis 1 (first part).

Role conflict as a moderator. With the motivation dependent variables, a bifurcation takes place; note that consideration and lack of role conflict are positively related ($r = .48$, $p < .001$). For high-role conflict, the consideration/motivation relationships remain high; however, for low-role conflict about half of the consideration/motivation relationships remain. For the behavior dimensions, those with moderate conflict have stronger consideration/behavior relations. The findings for high conflict are contrary to Hypothesis 2 (first part).

Internal-external control as a moderator. Here the results for the motivation data are much clearer. The relationships between consideration and motivation remain high for the internals; they become weak for the externals; so, in addition to internals having higher motivation than do externals, they also have a stronger relationship between supervisory consideration and motivation than externals. There is little impact on the consideration/performance relationship. Hypothesis 2 (second part) receives support.

These results suggest that internals have higher motivation and that their motivation is more predictable from leader behavior (consideration), that is, they perceive and respond to environmental contingencies more consistently than do externals. The impact of upward influences and role conflict is less clear-cut. However, when role conflict is high, supervisory consideration and motivation relations remain strong, the consideration/effort relationship is enhanced.

A further question can be raised concerning the leadership/motivation/behavior links: Can it be demonstrated that motivation occupies an intervening position between leadership and behavior as suggested in Hypothesis 3? One way to test this is to partial out motivation from the leadership/behavior relationship. The results for such a test are presented in Table 4. For the behaviors of effort, helping fellows and training subordinates, it appears that the partials drop substantially, suggesting that this linkage is appropriate; however, for the behaviors of doing what the boss tells or making suggestions to the boss, the consideration/behavior relationship remains just as strong after motivation has been partialled out.

The data for testing Hypothesis 4 (the relations between motivation and behavior) are presented in Table 4. For all behaviors except giving suggestions to boss, the hypothesis is supported (at $p < .05$). However, correlations are generally weak and motivation rarely explains more than 10% of the variance in behavior.

In Hypothesis 5 it is suggested that when motivation is higher, there is a stronger relation between behavior and satisfaction. In testing this hypothesis we noted an effect that confounded the contention that higher motivation results in a stronger relation between behavior and motivation. As motivation increases, both behavior and satisfaction increase concomitantly. The correlation data, taken alone, indicate that there is little relationship between effort and satisfaction under conditions of both high (−.13) and low (−.06) motivation, while the relation is strongest for the moderate motivation condition (.51, $p < .01$). However, the lack of relationship in the low- and high-motivation condition arises from different causes. An examination of the scatter diagrams indicates: (*a*) for low motivation there exists a distribution of effort and satisfaction over the whole of both scales; however, there is no relationship between the variables; (*b*) for high motivation the range of both scales is restricted; the cases cluster at the top of both.

DISCUSSION

These results provide an extension of the path-goal theories. Wahba and House (1974) have noted the strong rationality assumptions underlying the path-goal model. Their critique is given some indirect support in this study by noting the differences between internal subjects and external subjects in the applicability of the model to their behavior. Internal subjects who perceive their environment as meaningful and consistent and who feel able to control it are more likely to be rational decision makers in the sense proposed by economists and game theorists; while external subjects who see the environment as a cluster of random forces which they cannot control (even if it is benign) are more likely to make decisions that violate the rational assumptions of transitivity, independence of irrelevant alternatives, and dominance of alternative strategies. The results indicate stronger support for the model among the internal subjects than among the external subjects; this finding is consistent with the model's basis in rationality assumptions. Path-goal theorists, in trying to recreate a "goal-directed man" model of behavior, have to meet the rationality assumptions required. Their data have not yet done this.

There is a further problem which path-goal theorists have not yet adequately dealt with, that is, the question of conflicting path-goal instrumentalities. If an activity leads toward one cluster of goals and away from an equally important second cluster then, by our computations, the valence of that activity is neutral—neither desired nor not desired—in fact, the same as if the goals were unimportant or the activity led nowhere. However, Dalton (1947) has demonstrated that the group who were neither rate-busters nor quota-restricters were conflicted in the valence that productivity had for them; they were ambivalent toward it—they had intermediate performance and, it appears, they suffered from psychosomatic disorders. The path-goal theory can not yet predict these kinds of outcomes.

REFERENCES

BRAYFIELD, A. H., & ROTHE, H. F. An index of job satisfaction. *Journal of Applied Psychology*, 1951, **35**, 307–311.

DALTON, M. Worker response and social background. *Journal of Political Economy*, 1947, **55**, 323–332.

EVANS, M. G. Leadership and motivation: A core concept. *Academy of Management Journal*, 1970, **13**, 91–102. (a)

EVANS, M. G. The effects of supervisory behavior on the path-goal relationship. *Organizational Behavior and Human Performance*, 1970, **5**, 277–298. (b)

GALBRAITH, J. Path-goal models as a basis for the design of organizational reward and penalty systems. In B. S. Georgopoulos (Chm.), The path-goal relationship as a link between job performance and job satisfaction. Symposium presented at the meeting of the American Psychological Association, San Francisco, September 1968.

GAVIN, J. F. *Ability, effort and role perception as antecedents of job performance.* (Experimental Publication System MS No. 190A) Washington, D.C.: American Psychological Association, 1970.

GEORGOPOULOS, B. S., MAHONEY, G. M., & JONES, N. W. A path-goal approach to productivity. *Journal of Applied Psychology*, 1957, **41**, 345–353.

GRAEN, G. Instrumentality theory of work motivation: Some experimental results and suggested modifications. *Journal of Applied Psychology*, 1969, **53**(2, Pt. 2).

HOUSE, R. J. A path-goal theory of leadership. *Administrative Science Quarterly*, 1971, **16**, 321–338.

LAWLER, E. E. A correlational-causal analysis of the relationship between expectancy attitudes and job performance. *Journal of Applied Psychology*, 1968, **52**, 462–468.

LEWIN, K. The conceptual representation and the measurement of psychological forces. *Contributions to Psychological Theory*. Durham, N.C.: Duke University Press, 1938, **1**, No. 4.

MACDONALD, A. P. Internal-External Locus of Control. (Annotated Bibliography). Unpublished manuscript, West Virginia University, 1970.

PELZ, D. C. Influence: A key to effective leadership in the first-line supervisor. *Personnel*, 1952, **29**, 209–217.

PORTER, L. W. A study of perceived need satisfaction in bottom and middle management jobs. *Journal of Applied Psychology*, 1961, **45**, 1–10.

PORTER, L. W., & LAWLER, E. E. *Managerial attitude and performance*. Homewood, Ill.: Irwin-Dorsey, 1968.

RIZZO, J. R., HOUSE, R. J., & LIRTZMAN, S. I. Role conflict and ambiguity in complex organizations. *Administrative Science Quarterly*, 1970, **15**, 150–163.

ROTTER, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 1966, **80**(1, Whole No. 609).

SCHUSTER, J. R., CLARK, B., & ROGERS, M. Testing portions of the Porter and Lawler model regarding the motivational role of pay. *Journal of Applied Psychology*, 1971, **55**, 187–195.

STOGDILL, R. M. *Manual for the Leader Behavior Description Questionnaire—Form XII*. Columbus, Ohio: Bureau of Business Research, Ohio State University, 1963.

TOLMAN, E. C. *Purposive behavior in animals and men*. New York: Century, 1932.

VROOM, V. H. *Work and motivation*. New York: Wiley, 1964.

WARR, P. B., & ROUTLIDGE, T. An opinion scale for the study of manager's job satisfaction. *Occupational Psychology*, 1969, **43**, 95–109.

WAHBA, M. A., & HOUSE, R. J. Expectancy theory in work and motivation: Some logical and methodological issues. *Human Relations*, 1974, in press.

(Received February 12, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 2, 179-186

RACE, ECONOMIC CLASS, AND THE INTENTION TO WORK:

SOME NORMATIVE AND ATTITUDINAL CORRELATES[1]

JACK FELDMAN[2]

University of Florida

Four hypotheses, derived from a combination of instrumentality-expectancy theory and the concept of a nonresponsive environment for the hard-core unemployed, were tested. It was expected that (*a*) the hard core would perceive weaker normative pressures to seek work; (*b*) perceived normative pressure would have lower correlations with intentions to seek work in the hard-core samples than in the working-class samples; (*c*) the attitude toward each of five specific job-seeking behaviors would independently predict intended frequency of performance of each behavior; and (*d*) an instrumentality approach would predict intended frequency of each behavior less well in the black and the hard-core samples than in the white and the working-class groups. Results partially supported Hypotheses 1 and 2; partially reversed Hypothesis 3; and did not support Hypothesis 4. Data were interpreted as supporting the concept of white working-class alienation.

The general instrumentality-expectancy model (Fishbein, 1967; Graen, 1969; Porter & Lawler, 1968; Vroom, 1964) is generally applied to job performance and satisfaction, as well as to the broader issue of the attitude-behavior relationship. It can also be a fruitful source of hypotheses about the job-seeking behavior of different race and social class groups.

For example, Graen (1969) has shown that his model holds only in a "reciprocating climate" where performance is regularly evaluated, good performance rewarded, and contingencies clearly communicated. The job-seeking environment of the hard-core unemployed, especially blacks, can be considered nonreciprocating because of a lack of job-relevant skills and, in the case of blacks, racial discrimination. Thus, it can be predicted that the instrumentality model will not apply to these populations (for a more complete discussion see Feldman, 1972b).

Furthermore, the hard core cannot attain many socially desirable goals. Under these conditions, Cohen (1955) predicted the development of norms contradictory to those of the larger society. The work-related norms of the black groups or the hard-core groups, then, should differ in direction, strength, or both from those of the white or the working-class groups.

Some theoretical issues are also relevant to the study of job-seeking behavior. Fishbein (1967, chap. 51) has proposed and successfully demonstrated (Ajzen & Fishbein, 1970) that the overall attitude (defined in terms of pure affect or evaluation) toward the object (o) of an act (Ao) generally does not directly predict a specific behavior toward that object. Rather, the attitude toward the specific act (A act),[3] combined with normative beliefs relevant to the action, predicts the individual's behavioral intention with respect to the particular act. These intentions mediate between the model and measures of overt behavior and are thus useful variables to study in their own right (Ajzen & Fishbein, 1969).

---

[1] This study is based on a dissertation submitted in partial fulfillment of the requirements for the PhD degree at the University of Illinois. This research was supported by the Social and Rehabilitation Service of the Department of Health, Education, and Welfare, Grant 15-p-55175/5, Harry C. Triandis, principal investigator.

[2] The author wishes to thank Harry C. Triandis, Roy S. Malpass, Jim Jaccard, Andy Davidson, and Walter Hill for their comments on an earlier version of the manuscript and Joe McGrath, George Graen, Charles Hulin, and Martin Wagner for their help as members of the author's dissertation committee.

Requests for reprints should be sent to Jack M. Feldman, Department of Management, University of Florida, Gainesville, Florida 32611.

[3] The evaluation of the act (A act) is defined as the subjective expected utility of the act, a function of the act's expected consequences and their evaluation.

Joint consideration of Fishbein's (1967) theory and Graen's (1969, p. 22) revised model brings to mind the question of whether the intentions to use specific job-seeking behaviors are related to the attitude toward holding a steady job (Ao), which is weighted by the expectancy that the behavior leads to a job (Graen's "attitude toward the act as a means of attaining the role"), or to the attitude toward the act itself, which is determined entirely on the basis of its intrinsic outcomes (A act or Graen's "internal pressure").[4]

Modifications of the models cited above allow some variables relevant to job-seeking intentions to be described. To conform to previous work, the variables are stated in the form of a multiple regression equation.

Behavioral intentions (BI) are defined as the intended frequency of performance of given specific job-seeking behaviors.[5] Intentions are predicted by the following: (*a*) Two estimates of attitude toward working at a steady job (Ao): These are computed by taking the sum of products of measures of the perceived instrumentality of working for, and the corresponding evaluations of, two sets of work outcomes. The first is a standard set of outcomes (described later) and is called Ao_{JDI}. The second is based on outcomes elicited from members of the populations of interest, and is called Ao_{pop}. (*b*) Two estimates of normative pressure to seek work: The first is the sum of products of ratings of perceived direction and pressure (from four sources) of a general norm to "seek work if unemployed." The second is the sum of products of ratings of the same variables for a norm specific to each behavior. These are called NP general and NP specific, respectively. (*c*) An estimate of A act (or "internal pressure"): The sum of products of the evaluation of the intrinsic outcomes of each job-seeking behavior times the expectancy that the outcome is received is the estimate of A act. (*d*) The individual's directly measured evaluation of working times the expectancy that a particular behavior leads to a job is symbolized by (VwPw).

In regression equation form, this becomes:

$$\begin{aligned} BI = {} & (Ao_{JDI})w_0 \\ & + (Ao_{pop})w_1 + (\text{NP general})w_2 \\ & + (\text{NP specific})w_3 + (\text{A act})w_4 \\ & \qquad + (VwPw)w_5 + a \end{aligned}$$

The following hypotheses relate these variables to the environments of different socioeconomic groups.

Hypotheses

1. The hard core of both races perceive weaker pressure from friends, family, and self to get a job if unemployed than do the working class. Perception of general norm direction is also weaker for the hard core.
2. Both general and specific normative terms predict behavioral intentions less strongly in the hard core than in the working-class samples.
3. Following Fishbein, neither Ao_{JDI} nor Ao_{pop} independently predict behavioral intentions; A act (or "internal pressure") independently predicts BI.
4. Multiple correlations of the variables with BI in the black and the hard-core groups are lower than those in the white and the working-class samples.

METHOD

Subjects

Subjects were 100 black and 100 white males, between the ages of 18 and 50, living in the St. Louis, Missouri, metropolitan area. They were paid volunteers, recruited from social-service agencies (in the case of the hard core) or from local business and industrial firms (in the case of the working class). *Hard core* subjects were defined as those having been unemployed for at least three months prior to contact and being defined by a social-service agency as having problems in obtaining and keeping jobs. Working class subjects were defined as those who were currently employed at a full-time job (at least 35 hours per week) and who lacked known employment problems.

---

[4] It should be noted that the definition of (Ao) and (A act) are based on Graen (1969), though the terminology, for convenience, is Fishbein's. Some question exists as to the conceptual equivalence of the Graen and Fishbein formulations of these terms, though the author considers them to be equivalent. Graen's theory uses the difference between the expectancy that "standard" effort leads to role attainment and the expectancy that "superior" effort leads to role attainment. Since multiple, nonmutually exclusive behaviors are considered in this study, only a single expectancy term is used.

[5] This measure was chosen because it seemed to be the closest analogy to effort expended on the job, a variable of interest in other studies.

The working-class sample excluded professionals and managers beyond first-line supervisors but included some white-collar employees. The majority of the sample were skilled and semiskilled blue-collar workers (for details, see Feldman, 1972a, 1972b). About 50 subjects fell in each cell of a 2 × 2 sampling design (black versus white, hard core versus working class).

Of course, this sample was somewhat biased. It included, for example, no extremely militant blacks, since they refused to participate. There also were many hard-core unemployed who did not come to the attention of social-service agencies for various reasons. There were probably few alcoholics, drug addicts, or people with serious criminal records in the sample. Also, the use of volunteers may have introduced bias. Thus, generalizability is somewhat limited.

Interviewers

Interviewers were young (20-35 years old) black men and white men, recruited locally. All were high school graduates, either working or attending college, and were selected by the author or his associate (the director of Psychological Services, Narcotics Service Council, a private drug rehabilitation agency).[6] All interviewers were initially trained by the author; some replacements were trained by the field supervisor who had special preparation for such contingencies.

Data were collected as part of a larger interview during a period of two to four hours (see Feldman, 1972a). Subjects were paid $8 and interviewers $10 for a completed questionnaire. All interviews were arranged at a mutually convenient time and place.

Instruments and Procedure

The direction and intensity of a subject's evaluations of 15 job outcomes was measured by choosing one of the nine faces of the Kunin (1955) scale presented on a flashcard. The first 5 standard outcomes were the paraphrased factor names of the Job Descriptive Index (JDI) (Smith, Kendall, & Hulin, 1969): good pay (pay), working with people you like (co-workers), having a good boss (supervisor), being promoted (promotion), and enjoying the work you do on the job (work itself). The next 10 outcomes, elicited from a sample of black working class, white working class, and unemployed men (for details see Triandis, Feldman, & Harvey, 1971), were the most frequent responses to the question: "Name three things you feel you get from working (at a regular job)." These were the following: having responsibilities, owing money, saving money, buying nice things, being bored, family's respect, friend's respect, having fun, being tired, and supporting yourself.

Five specific job-seeking behaviors were used: going to an employment agency, going from place to place filling out applications, trying to join a training program, reading want ads, and hanging around with friends waiting for something to turn up. These were frequently mentioned by a sample of blacks and of whites as typical job-seeking methods.

Five intrinsic outcomes of each job-seeking behavior, defined as "things that happen to you if you use that particular method," were also elicited. These are omitted for reasons of space but can be found in Feldman (1972b) or obtained from the author.

The evaluations of working at a regular job (more than 35 hours per week), of each of the five job-seeking behaviors, and of the five intrinsic outcomes of each behavior were also measured by Kunin's scale.

Norm direction was measured by a 3-point scale (3 = should, 2 = may or may not, 1 = should not) presented on a flashcard. Norm pressure was measured similarly (strongly insist = 3, care a little = 2, don't care at all = 1). The direction and pressure of the general norm, and a specific norm for each job-seeking behavior, were measured for four sources: friends, family, people in general, and self.

Instrumentality, the perceived correlation between working and positive and negative outcomes, was measured by a 3-point scale (1, 0, −1) and again presented on a flashcard. Expectancy, the perceived probability of an outcome given the performance of a job-seeking behavior, was measured by asking subjects to estimate the chances in 10 of the outcome occurring, as if they were making a bet. The perceived probability that a given job-seeking behavior leads to finding a job (Pw) was measured by this method also. Behavioral intentions were measured on a 5-point scale (0 = never, 4 = every day) presented on a flashcard.

Procedure

All items and instructions were decentered by a modification of the Werner and Campbell (1970) technique. Black undergraduates, whose homes were in ghetto areas, translated a set of items into "black" English then retranslated a different set into "standard" English. This procedure was repeated until a single version, comprehensible to both blacks and whites, was obtained. This version was pretested by the interviewers on black subjects and white subjects from the populations of interest. Any language difficulties were resolved before data collection.

Practice sheets, simple examples of each response, were included before each rating task. This allowed a comprehension check and was also intended to reduce warm-up effects (for full details, see Tri-

---

[6] The author wishes to gratefully acknowledge the fine work of William Harvey and the staff at Nasco West in locating subjects and coordinating interviews.

TABLE 1

PERCEIVED NORMATIVE PRESSURE FROM FOUR SOURCES TO LOOK FOR WORK

Norm source	Economic class				Proportion of variance		
	Hard core		Working class				
	Race						
	Black	White	Black	White	R	E	I
Friends	2.42	2.27	2.50	2.31	—	—	—
Family	2.92	2.90	2.98	2.94	—	—	—
People in general[a]	1.94	2.10	2.23	1.69	—	—	.06
Self[b]	2.90	2.79	3.00	2.94	—	.04	—

Note. Ratings range from 1 (low) to 3 (high). All means are rounded to second decimal. Proportion of variance (univariate effects) refers to estimated omega-squared (Hays, 1963). R = race effect, E = economic-class effect, I = interaction effect.

[a] Significant univariate interaction effect ($p < .0003$).

[b] Significant univariate economic-class main effect ($p < .004$).

andis, 1972; for examples of practice items, see Feldman, 1972a).

Subjects were interviewed in a standard order: evaluation, instrumentality, expectancy, norms, behavioral intentions, employment history, and demographic data. Within a topic, order of methods was counterbalanced (for details of the multitrait, multimethod validation of evaluation and expectancy measures, see Feldman, 1972a). Subjects who could not comprehend a task (after re-instruction) at any point were excused with a small payment. Less than 9% of the black hard-core sample and less than 5% of the other samples had to be replaced due to inability to perform the tasks. No other subject attrition occurred.

Interviewers used a neutral tone and manner developed through role playing. To minimize biasing factors, interviewers were never told the specific hypotheses under investigation and were always of the subject's own race. In addition, 10% of the subjects were called at random and questioned about the interview to ensure the quality of the data.

RESULTS

Validation of Sampling Strategy

Analyses of a series of demographic and job-history questions revealed no unusual between-sample differences that might confuse interpretation of the data.[7] Since causality does not reside in race or social class, but in the experiences one has as a result of race or class, such data are potentially quite useful.

Tests of Hypotheses

Hypothesis 1, that the hard core of both races perceive weaker pressure to find work if unemployed and have weaker norm-direction perceptions as well, was partially supported. A significant multivariate economic-class effect ($F = 2.53$, $p < .05$) was found for ratings of perceived pressure to look for work. As Table 1 shows, this was due to a significant univariate effect, in the predicted direction, for pressure from one's self.

A significant multivariate interaction was also present ($F = 3.79$, $p < .006$). This was due to a univariate interaction on ratings of pressure from people in general. The black working class reported higher pressure from this source than any other group.

Table 2 illustrates similar results for ratings of norm direction. A significant multivariate economic-class effect ($F = 3.19$, $p < .02$) was due to the hard-core sample's tendency to make fewer "should do" responses for ratings of personal normative beliefs and those of "people in general."

A significant multivariate interaction also occurred ($F = 6.10$, $p < .0002$). Univariate tests show that it was based on the black hard core's tendency to perceive weaker norm-direction from "people in general" than that of the white hard core, while the opposite was true in the working class.

Hypothesis 2, that both general and specific normative terms (NP) have lower correlations with behavioral intentions in the hard-core samples, was not supported. As Table 3 shows, the correlations of general and specific NP with behavioral intentions were never significant for the black hard core, but were significant 7 of 10 possible times for the white hard core. Many behavior- and sample-specific effects were also present.

---

[7] The multivariate analysis of variance, which provides a simultaneous test of significance for all dependent variables, was used to analyze this data. Univariate analyses of each dependent variable were also conducted, to more precisely identify the source of any observed overall differences. Univariate effects not associated with significant multivariate effects were not reported, to minimize the possibility of a Type 1 error. Jones (in Cattell, 1966) provided a discussion of this technique.

Since race and socioeconomic class are correlated in the population, a purely random sampling design reflects this correlation. In order to meet the assumptions of analysis of variance, these variables must be made independent (unconfounded). The analyses of job-history and demographic variables show that this goal was, in fact, achieved (for a brief summary, see Feldman, 1973; full details are reported in Feldman, 1972a).

Hypothesis 3, that attitude toward working (Ao) does not predict behavioral intentions, but attitude toward the specific act (A act) independently predicts behavioral intentions, was not supported. Table 4 presents the standardized regression weights and multiple correlation for each behavior in each sample. A act never significantly predicted behavioral intentions, while Ao, based either on the Job Description Index or population-elicited outcomes, was significant 4 out of a possible 10 times for the black working class. Once again, many sample- and behavior-specific effects existed.

Table 4 also illustrates a failure to support Hypothesis 4, that the black and the hard-core samples' intentions are less predictable than those of the white and the working class. The white hard-core and the black working-class samples were the most predictable, on the basis of the number of significant multiple correlations. Two-way (Race × Economic Class) analyses of variance (Jones, 1968) were performed on the multiple correlations for each behavior. Only a very marginal ($\chi^2 = 2.95$, $p < .10$) interaction on intention to go to an employment agency and a somewhat stronger ($\chi^2 = 4.44$, $p < .05$) interaction on intention to read want ads were found. Neither was in the predicted direction.

TABLE 2

PERCEIVED NORM DIRECTION FROM FOUR SOURCES TO LOOK FOR WORK

Norm source	Economic class				Proportion of variance		
	Hard core		Working class				
	Race						
	Black	White	Black	White	R	E	I
Friends	2.65	2.77	2.79	2.73	—	—	—
Family	2.98	3.00	3.00	2.96	—	—	—
People in general[a,b]	2.40	2.71	2.54	2.17	—	.03	.09
Self[a]	2.98	3.00	3.00	2.98	—	.02	—

Note. Ratings range from 1 (should not) to 3 (should). All means rounded to second decimal. Proportion of variance (univariate effects) refers to estimated omega-squared (Hays, 1963). R = race effect, E = economic class effect, and I = interaction effect.

[a] Significant univariate economic class main effect ($p < .04$ or better).

[b] Significant univariate interaction effect ($p < .0001$).

TABLE 3

CORRELATION OF GENERAL AND SPECIFIC NORMATIVE PRESSURE WITH FIVE BEHAVIORAL INTENTIONS IN FOUR SAMPLES

Behavior	Norm pressure		
	Sample	General	Specific
Agency	BHC	−.07	−.04
	BWC	.17	.24
	WHC	.42***	.56****
	WWC	.27	.03
Applications	BHC	−.09	−.07
	BWC	.17	.22
	WHC	.44***	.46****
	WWC	.15	.35**
Training	BHC	.02	.10
	BWC	.43***	.41***
	WHC	.21	.19
	WWC	.31*	.09
Ads	BHC	−.13	.08
	BWC	.35**	.30*
	WHC	.52****	.50****
	WWC	.33**	.12
Friends	BHC	.00	.22
	BWC	−.06	.08
	WHC	−.40***	.19
	WWC	−.19	.19

Note. All figures rounded to two decimals. BHC = black hard core; BWC = black working class; WHC = white hard core; WWC = white working class. For BHC, $n = 48$; for BWC, $n = 52$; for WHC, $n = 52$; for WWC, $n = 48$.

* $p < .05$.

** $p < .02$.

*** $p < .01$.

**** $p < .001$.

DISCUSSION

Limitations on Internal and External Validity

The results were, for the most part, not in the predicted direction. It should be noted that response-bias effects are probably not responsible. Means and standard deviations for all variables in all samples were systematically compared, and no consistent pattern of extremely high or low means combined

TABLE 4

PREDICTION OF FIVE BEHAVIORAL INTENTIONS IN FOUR SAMPLES BY THE INSTRUMENTALITY MODEL

Behavior	Standardized regression weights							
	Sample	Ao_{JDI}	Ao_{pop}	$NP_{general}$	$NP_{specific}$	A act	VwPw	*R*
Agency	BHC	−.25	−.03	−.13	.04	.03	−.06	.29
	BWC	.13	.25	.08	.30	−.09	−.10	.38
	WHC	.04	.22	.22	.48****	.02	−.14	.62***
	WWC	.01	.01	.30	.05	−.16	.08	.32
Applications	BHC	.16	.10	−.04	−.09	.31	−.14	.37
	BWC	.15	42***	.08	.19	−.03	.07	.48
	WHC	.01	.10	.17	.20	.01	.29	.58***
	WWC	.18	.04	−.00	.32	.09	.07	.40
Training	BHC	−.12	−.06	−.21	.17	.12	.25	.35
	BWC	−.03	.41***	.44***	.02	−.02	.00	.59***
	WHC	−.25	.01	.09	.07	−.02	−.01	.32
	WWC	−.01	−.12	.31*	.06	.12	.12	.38
Ads	BHC	.00	−.05	−.29	.27	.29	−.12	.38
	BWC	.34**	.25*	.25	.15	.13	.02	.61***
	WHC	−.07	.20	.24	.19	−.03	.29*	.67***
	WWC	−.23	−.09	.33**	−.03	−.19	−.15	.46
Friends	BHC	.05	.01	.01	.21	.32	−.12	.37
	BWC	.08	.08	−.10	.06	.30	−.29	.37
	WHC	.04	.22	.32*	.09	.01	−.14	.50*
	WWC	−.08	−.01	−.15	.08	−.06	−.35*	.45

Note. All figures rounded to third decimal. Ao = attitude toward the object; A act = attitude toward the act; NP = normative pressure; Ao_{JDI} = outcomes from the Job Descriptive Index; Ao_{pop} = outcomes elicited from members of the populations of interest; VwPw = the individual's directly measured evaluation of working times the expectancy that a particular behavior leads to a job. For BHC, *n* = 48; for BWC, *n* = 52; for WHC *n* = 52; for WWC, *n* = 48.

* $p < .05$.
** $p < .025$.
*** $p < .01$.
**** $p < .001$.

with low standard deviations was found. Thus, response bias in one sample or the other seems unlikely. It is simpler to assume that the observed differences are valid than to propose response biases which operate in different directions for different variables, and sometimes not at all.

Interviewer bias or demand-characteristic effects were also possible. However, since interviewers were never told the exact hypotheses, and since some involve combinations of variables and the prediction of correlations rather than mean differences, it seems unlikely that the obtained results were due to bias. Furthermore, the interview was highly standardized, more like an oral questionnaire; interviewers were trained and tested in the use of neutral tones of voice and expression; they were also limited in the interpretations they could provide. The importance of unbiased data was stressed repeatedly, and informal observation and discussion indicated strong motivation to perform properly. The above precautions do not rule out biased data, but such a possibility is less likely than it would be otherwise.

Race and Economic-Class Differences

It seems true that the hard core of both races feel less normative pressure to look for work, at least in terms of their "own" norm. The black working class appears to feel normative pressure from self and society most strongly. The differences were not large, but this may have been due to a lack of response options.

However, only in the black hard-core sample was there no relationship between norms and behavioral intentions. This finding contrasts sharply with the white hard core to when norms seem important. It is tempting to conclude that ideas of what one "should do" are not strong determinants of the black hard core's behavior, but such a conclusion is premature. Besides the inability to infer causality from correlational data and the need for cross-validation in any case, it may be that the norm sources used were not relevant to the black hard core. The hypothesis is reasonable, however, and deserves experimental investigation.

The failure to support Hypothesis 3 may indicate important group differences. Fishbein (1967) has stated that attitude toward the act and/or normative beliefs, but *not* attitude toward the object, should independently predict behavioral intentions. This is usually found to hold true. Job seeking, however, may be a special case, since the loss of a job can be a personal disaster. Any behavior which might lead to a job, whatever the probability, may be performed. Thus, the more one values work, the more one should look for a job, regardless of the evaluation of the behaviors involved.

This attitude toward work and job seeking seems to be the case in the black working class. Why it is not in either hard-core sample is suggested by correlational evidence (reported more fully in Feldman, 1972b). In the combined hard-core sample, evaluation of work was significantly negatively correlated with the skill level of the last job held and whether nonpublic aid income was received. Two behavioral intentions were also negatively correlated with the receipt of nonpublic aid money.[8] The black working class may regard work as a source of outcomes beside money, and the hard-core samples may not. For the hard core, any source of income might suffice. Thus, only normative pressure predicted behavioral intentions for the white hard core (with one exception), while the black hard core were not predictable at all.

TABLE 5

INTENDED FREQUENCY OF PERFORMANCE OF SIX JOB-SEEKING BEHAVIORS

Behavior	Economic class				Proportion of variance		
	Hard core		Working class				
	Race						
	Black	White	Black	White	R	E	I
Look for work[a]	3.69	3.42	3.58	3.25	.02	—	—
Go to an employment agency[a]	2.77	1.93	2.94	2.03	.10	—	—
Fill out many applications[a]	3.23	3.13	3.29	2.65	.02	—	—
Read want ads[b]	3.23	3.13	3.48	3.50	—	.02	—
Try to join a training program[a, b]	1.17	.42	1.77	.96	.11	.05	—
Hang around with friends[c]	1.60	2.02	1.90	1.29	—	—	.02

Note. Multivariate effects: race ($F = 7.176$, $p < .0001$), economic class ($F = 5.25$, $p < .0001$), interaction ($F = 2.40$, $p < .03$). All means are rounded to two decimals. Five-point scale was used: 0 = never, 4 = every day. Proportion of variance (univariate effects) refers to estimated omega-squared (Hays, 1963). R = race effect, E = economic-class effect, I = interaction effect.

[a] Univariate race effect ($p < .05$ or better).

[b] Univariate economic-class effect ($p < .05$ or better).

[c] Univariate interaction effect ($p < .02$).

The white working class, whose intentions were also unpredictable, may have a different orientation. If they regard work only as a way to support themselves, but have no other money-making skills, a regular job is vital. Thus, evaluation of work does not relate to intentions to seek it, because they perceive no options. Feldman (1973) has shown that black working-class men believe they can get along well if unemployed, but white working-class men do not. Therefore, the blacks' evaluation of work should correlate with job seeking, since they perceive a choice.

The fact that not all behaviors were predictable also calls for some comment. It may have been that some behaviors had very high- or low-base rates in some samples and these reduced the obtained correlations. This possibility is partially supported by Table which shows that the only unpredictable behavior for the white hard core has an extremely low intended frequency. It does no completely explain the failure to predict thre of five behaviors in the black working clas since "want ads" has a fairly high intende frequency. Also interesting is the finding tha training programs appear much more salien for black than for white subjects.

The fact that the attitudinal terms pr dicted behavioral intentions for the blac working class and normative for the whi hard core, while the white working class and the black hard core were not predictable at all, suggests that recent social and economic reforms have not yet reached the black hard core and have negatively influenced the white working class. Changing these perceptions and attitudes is not a psychological, but a social problem. They will change when opportunities for rewarding employment change, and not before.

---

[8] The correlations of these and other job-history and demographic variables with evaluation and intentions measures are reported fully in Feldman (1972a, 1972b).

REFERENCES

AJZEN, I., & FISHBEIN, M. The prediction of behavioral intentions in a choice situation. *Journal of Experimental Social Psychology*, 1969, **5**, 400–416.

AJZEN, I., & FISHBEIN, M. Attitudinal and normative variables as predictors of specific behaviors: A review of research generated by a theoretical model. Paper presented at the Workshop on Attitude Research and Consumer Behavior, University of Illinois, December 1970.

COHEN, A. K. *Delinquent boys: The culture of the gang*. Glencoe, Ill.: The Free Press, 1955.

FELDMAN, J. M. Race, economic class, and job-seeking behavior: An exploratory study. Unpublished doctoral dissertation, University of Illinois, 1972. (a)

FELDMAN, J. M. Race, economic class, and job-seeking behavior: An exploratory study. (Report 15, SRS No. 15-p-55175/5) Champaign, Ill.: University of Illinois, Department of Psychology, 1972. (b)

FELDMAN, J. M. Race, economic class, and perceived outcomes of work and unemployment. *Journal of Applied Psychology*, 1973, **58**, 16–22.

FISHBEIN, M. Attitude and the prediction of behavior. In M. Fishbein (Ed.), *Readings in attitude theory and measurement*. New York: Wiley, 1967.

GRAEN, G. Instrumentality theory of work motivation: Some experimental results and suggested modifications. *Journal of Applied Psychology*, 1969, **53**(2, Pt. 2).

HAYS, W. L. *Statistics for psychologists*. New York: Holt, Rinehart & Winston, 1963.

JONES, L. Analysis of variance in its multivariate developments. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology*. Chicago: Rand McNally, 1966.

JONES, M. B. Correlation as a dependent variable. *Psychological Bulletin*, 1968, **69**, 41–55.

KUNIN, T. The development of a new type of attitude measure. *Personnel Psychology*, 1955, **8**, 65–78.

PORTER, L. W., & LAWLER, E. E. *Managerial attitudes and performance*. Homewood, Ill.: Irwin, 1968.

SMITH, P. C., KENDALL, L. M., & HULIN, C. L. *Measurement of satisfaction in work and retirement*. New York: Rand McNally, 1969.

TRIANDIS, H. C. *The analysis of subjective culture*. New York: Wiley-Interscience, 1972.

TRIANDIS, H. C., FELDMAN, J. M., & HARVEY, W. M. The perception of implicative relationships among black and white adolescents and the hardcore unemployed. (Report 9, SRS No. 15-p-55175/5. Champaign, Ill.: University of Illinois, Department of Psychology, 1971.

VROOM, V. *Work and motivation*. New York: Wiley, 1964.

WERNER, O., & CAMPBELL, D. T. Translating, working through interpreters, and the problem of decentering. In R. Narrol & R. Cohen (Eds.), *A handbook of method in cultural anthropology*. New York: American Museum of Natural History, 1970.

(Received January 22, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 2, 187–191

IMPROVING JOB PERFORMANCE THROUGH TRAINING IN GOAL SETTING[1]

GARY P. LATHAM[2]
University of Akron

SYDNEY B. KINNE III
Georgia Kraft Company, Rome, Georgia

The present study examined the generality of goal-setting theory to an industrial setting. Twenty pulpwood-logging operations were matched and randomly assigned to either a one-day training program in goal setting or a control group. Measures of production, turnover, absenteeism, and injuries were collected for 12 consecutive weeks. Analyses of variance indicated that goal setting can lead to an increase in production and a decrease in absenteeism.

Research on goal-setting theory (Locke, 1968) has been criticized for its lack of emphasis on investigating employee behavior in a natural work environment (Heneman & Schwab, 1972). Data supporting the theory are based largely on laboratory experiments in which the dependent variable is typically a task requiring simple addition. As Campbell, Dunnette, Lawler, and Weick (1970) have stated, the differences between college students solving addition problems and the behavior of workers in industrial settings must be considered.

Criticism on the theory's lack of external validity is not completely justified. Latham and Ronan (1970) factor analyzed data obtained from a random sample of 292 pulpwood producers. The results showed that goal setting and supervision loaded on the same factor as two measures of production (positive) and one measure of injuries (negative). Goal setting without supervision loaded on the only factor containing measures of turnover (positive). Supervision that did not include goal setting loaded on a third factor that did not include any criterion variables. No relationship was found between goal setting or supervision and absenteeism.

Although zero-order correlations between the previously mentioned independent and dependent variables were generally significant at the .05 level, the values were quite low. In order to corroborate these results, Latham and Kinne (1971) collected data on 892 additional producers and conducted an analysis of variance on the man-hour production. The results indicated that producers who supervise their employees and set production goals have higher productivity than producers who supervise their men but do not set production goals. In summarizing the results of these two studies, Ronan, Latham, and Kinne (1973) interpreted their findings as supporting Locke's (1966) contention that supervision leads to superior performance only insofar as it results in the establishment of specific performance goals. However, supervision, as defined by staying on the job with the crew, is required to insure that the worker accepts these goals. Goal setting without the presence of supervision may lead to an increase in labor turnover.

A limitation of the above studies is that both were correlational in nature. In no instance was the independent variable manipulated by the authors. The purpose of the present study was to determine the effects of a one-day training program on goal setting on the job performance of pulpwood workers. The primary criteria for measuring the effectiveness of the program were cords-per-sawhand-hour and cords-per-crew-hour production. Additional criteria included measures of turnover, absenteeism, and injuries.

---

[1] This study is based in part on a technical report written by the authors for the American Pulpwood Association, Harvesting Research Project. The opinions expressed herein are those of the authors and do not necessarily reflect the position of the American Pulpwood Association or the Georgia Kraft Company.

The authors are grateful to Edwin A. Locke and Gary A. Yukl for reading and commenting on an earlier draft of this manuscript.

[2] Requests for reprints should be sent to Gary P. Latham, who is now at the Department of Human Resource Planning and Research, Weyerhaeuser Company, Tacoma, Washington 98401.

METHOD

Sample

Twenty-six producers and their crews were selected and matched on the following variables: (*a*) crew size, (*b*) cords-per-man-hour production, (*c*) geographic location, (*d*) logging system, (*e*) delivery point, (*f*) sawhand function, (*g*) delivery system, (*h*) number of hours the producer spent on the job site with the crew, and (*i*) "did not previously set a daily, weekly, or monthly production goal." One crew from each pair was randomly assigned to the training group; the remaining subjects constituted a control group. The method of payment for both groups was their normal piece-rate system. No monetary or nonmonetary bonus other than verbal praise was tied to goal attainment.

Procedure

Seven company foresters who were aware of the aims and objectives of the producer's job, who frequently observed the producer on the job site and who were well-known by the producers, were given instruction in training. A three-hour program was conducted in which the purpose of the experiment and the assumptions underlying it were explained. This program was followed by an intensive discussion period.

Each forester contacted the selected producers and requested their participation in the study. The explanation given to the experimental and control groups was similar: Each was told that they would be participating in a research program, that the study would last three months, that they were under no obligation to enter the program, and that they could leave the program at any time. The essential difference in the instructions to the two groups was that the experimental subjects were told that the purpose of the study was to give them a one-day training program in learning to set production goals and to determine its effects on performance. The control subjects were told that the study was designed to determine the effects of injuries, absenteeism, and turnover on production.

Three producers assigned to the experimental group expressed an unwillingness to participate in the study. This nonparticipation necessitated the removal of the 3 control producers who had been paired with them. Thus, the results of this study are based on 20 producers.

Training in goal setting was based on the premise that increasing the performance of the sawhand results in an increase in the productivity of the crew. This premise was based on the fact that the felling of a tree is vital to achieving the remaining tasks in pulpwood harvesting. By using regression equations that had been developed from actual field measurements to predict cords per man-hour and number of trees cut per sawhand-hour, production tables were constructed to systematically establish production goals.[3] These tables took into account factors that could affect production due to differences in harvesting stand conditions.

On the first working day of each week, the trainer visited the experimental and control producers on the job site. The average diameter breast height based on 30 to 40 trees to the nearest inch was determined for the harvesting area. By using the table which represented the sawhand function and stand condition, a production goal in trees per sawhand hour was determined for the existing day and/or week. This goal was explained to the producers in the experimental group who in turn explained it to their sawyers. This goal was stressed as a minimum standard of acceptable performance; however, no penalties were provided for failure to attain the goal. The production goal was not made known to producers in the control group.

Tally meters were provided for the sawyers in the experimental group so that they could constantly monitor or evaluate their performance. The assumption at this point was that each number accumulated on the tally meter provides information feedback to the employee and thus is reinforcing for the individual. The possible confounding effects of goal feedback with goal setting were considered minimal, since previous research has shown that knowledge of score alone does not lead to higher performance unless it is used to set goals (Locke, Cartledge, & Knerr, 1970; Locke, Cartledge, & Koeppel, 1968).

Producers in both the experimental and control groups were given a form similar to that used by Latham (1971) to record production, turnover, absenteeism, and injuries. The only difference between the form given to the experimental group and that given to the control group was that the former contained an item requesting information concerning the production goal. The trainer had a questionnaire on which the production goal for the control group was recorded. These forms served the same purpose as the tally meter for the sawyer, namely, knowledge of results to the producer.

Data were collected for 14 consecutive weeks. The three-month study period was employed in order to determine the effects of the one-day training program over a considerable length of time. The results of only 12 weeks were retained, inasmuch as mill shutdowns occurred in 2 of these weeks, and hence the majority of the sample were unable to deliver timber.

RESULTS

The results of this study were based on six performance scores. The two primary criteria, cords per sawhand-hour and cords per crew-hour, were based on the arithmetic difference between actual production and expected (goal) production for the sawhand and the crew. Expected production was the assigned performance goal determined from the production tables.

---

[3] These equations were developed by Georgia Kraft Company prior to this study. Time studies were made of more than 300 logging operations on varying harvesting conditions. The resulting production goal may be characterized as moderate in difficulty in that average to above average production is necessary in order to achieve the goal. Information concerning these tables can be obtained from the second author.

The four remaining criteria included measures of turnover, absenteeism, and injuries. Turnover was defined by two measures, namely, the number of men who quit and the number of men who were fired. Each of these variables was divided by the total number of men in the producer's work force. Injury rate was defined as the number of men hurt on the job who missed eight or more hours of work. Absenteeism was defined as the number of men off the job eight or more hours for reasons other than an injury. Both of these variables were divided by the number of men in the crew. These definitions are identical to those used in a previous study by Latham (1971).

An analysis of variance was conducted to determine the significance of goal setting and weeks on the production of the individual sawhands. The difference in production between the group of sawyers who received training in goal setting and the control group was in the expected direction but the difference was only marginally significant ($F = 3.43$, $df = 1/9$, $p < .10$). The performance means and standard deviations of the individual sawyers in the experimental and control groups are shown in Table 1. The effect of weeks and the Weeks × Treatment interaction (goal training) was not significant.

TABLE 1

SUMMARY OF DIFFERENCES BETWEEN GOAL AND ACTUAL PERFORMANCE FOR SAWHAND IN CORDS PER SAWHAND-HOUR

Producer pair	Treated		Control	
	$\bar{X}$	*SD*	$\bar{X}$	*SD*
1	2.05	.43	.58	.83
2	−.25	.58	−.11	.69
3	.20	.56	−.17	.48
4	.02	.79	−.13	.73
5	.20	.45	.05	.61
6	.64	.97	−.09	.43
7	.36	.58	−.45	.59
8	1.26	.63	1.68	.68
9	.69	.88	−.21	.93
10	.44	.35	.86	.57

TABLE 2

SUMMARY OF DIFFERENCES BETWEEN GOAL AND ACTUAL PERFORMANCE FOR CREWS IN CORDS PER MAN-HOUR

Producer pair	Treated		Control	
	$\bar{X}$	*SD*	$\bar{X}$	*SD*
1	.23	.04	.09	.15
2	−.09	.21	−.09	.29
3	.05	.20	−.08	.17
4	.05	.20	−.07	.20
5	.04	.09	.01	.11
6	.13	.20	−.02	.12
7	.08	−.14	−.16	.22
8	.22	.11	.24	.10
9	.08	.12	−.07	.29
10	.11	.07	.13	.12

A similar analysis was performed on the cords-per-man-hour production for the entire crew. First, this analysis provided a test of a key assumption in this study, namely, that increasing the productivity of the sawyer results in an increase in the productivity of the crew. In general, there is only one sawyer in a crew ranging in size from two to eight men. Second, the analysis provided a test of the balance of the producer's operation in terms of the number of his employees and his equipment. The results showed that the performance of the experimental crews was significantly higher than that of the control group ($F = 7.44$, $df = 1/9$, $p < .05$), thus supporting the assumption that the sawyer has a key position in the producer's work force. Again, the effects of weeks and the Weeks × Goal Setting interaction were not significant. Thus, it may be concluded that goal setting leads to a significant increase in the productivity of the group as well as the individual. Table 2 shows the means and standard deviations for each crew.

The finding that the difference between the experimental and control crews was at a higher level of significance than that between experimental and control sawyers is noteworthy. This result underscores the importance of a well-balanced harvesting system in terms of manpower and equipment. Any improvement in performance in one function in harvesting, even if only slight, can have an appreciable effect on the total performance of the operation when its effects are distributed (and hence magnified) over a well-balanced operation. Conversely, if performance in one function is greatly improved, its effects on the entire system are nullified to the extent that the operation is unbalanced.

Two-way analyses of variance (tables are not shown) were conducted to determine the effects of a month (four weeks) and goal setting on turnover, absenteeism, and injuries. No significant differences were found between months. Moreover, goal setting had no significant effect on the two measures of turnover or injuries. The number of people who quit or were fired or injured was very low in both groups. Absenteeism, however, was significantly higher ($p < .05$) in the control group than in the experimental group.

DISCUSSION

The present study has shown that training in goal setting can lead to an increase in production and a decrease in absenteeism. The theoretical advantage of approaching the problem of worker motivation through goal setting is that it is not dependent on the use of mythological terms such as id, ego, and superego; nor does goal setting postulate personality mechanisms, or drive states, or separate and distinct factors independent of each other that contribute differentially to short-term and long-term performance. However, goal setting, as a theory of motivation, does focus on variables intrinsic to the job, namely, the job task(s) itself. It is based on a learning model in that the worker must be taught to set a task objective and he should be given information feedback concerning his performance. Knowledge of results provides meaning to a task. It is probable that tree counting in the present study enabled the sawyer to obtain a sense of achievement. The number of trees cut assumed meaning in that the worker was able to determine the extent to which his performance was above average. It may have been this "meaningful" dimension that contributed to the lower absenteeism in the goal-training group, that is, goal setting led to effective performance, effective performance led to job satisfaction, and job satisfaction led to a reduction in absenteeism.

The use of the production tables as a systematic method for setting production goals is of theoretical importance. Work on levels of aspiration has shown that if the individual sets a performance goal before carrying out a task, he tends to raise the goal if he is successful, since he increases his expectations of himself (Lewin, 1951). Thus, to improve one's performance one must first aspire but to aspire one must see that success is possible, that is, clear evidence must be available that others under similar conditions are succeeding. The production tables were based on data collected from a large number of sawyers and hence provided this evidence. The sawyer knows that the goal is attainable and that it is not based on the whim of a demanding supervisor or on the unrealistic expectations that he himself has.

While there is an obvious need for additional research on the validity of goal-setting theory, the situation is by no means as serious as previous researchers have led us to believe. As a theory of motivation, goal setting appears to be effective in an industrial setting, and its effects appear to hold over time. A possible limitation of training in goal setting is that it is too simplistic for a complex situation, that is, it leads to a significant improvement in the performance of the average or above average worker, but it does little for the worker who is badly in need of a comprehensive training program. Nevertheless, even in such instances the supervisor who sets a specific task goal for his subordinates is in effect making it clear what it is they are supposed to do (Campbell et al., 1970). Thus, in the process of discussing with the worker the nature of these goals, the worker may acquire specific knowledge concerning his job tasks, their priorities, and the most effective methods that lead to their attainment.

REFERENCES

CAMPBELL, J. P., DUNNETTE, M. D., LAWLER, E. E., & WEICK, K. E. *Managerial behavior, performance, and effectiveness.* New York: McGraw-Hill, 1970.

HENEMAN, H. G., III, & SCHWAB, D. P. Evaluation of research on expectancy theory predictions of employee performance. *Psychological Bulletin*, 1972, 78, 1–9.

LATHAM, G. P. *Indicators of productivity, turnover, absenteeism, and injuries*. Atlanta, Ga.: American Pulpwood Association, Harvesting Research Project, 1971.

LATHAM, G. P., & KINNE, S. B., III. *Goal setting as a means of increasing the performance of the pulpwood harvester*. Atlanta, Ga.: American Pulpwood Association, Harvesting Research Project, 1971.

LATHAM, G. P., & RONAN, W. W. *The effects of goal setting and supervision on the motivation of pulpwood workers*. Atlanta, Ga.: American Pulpwood Association, Harvesting Research Project, 1970.

LEWIN, K. *Field theory in social science: Selected theoretical paper*. New York: Harper, 1951.

LOCKE, E. A. The relationship of intentions to level of performance. *Journal of Applied Psychology*, 1966, **50**, 60–66.

LOCKE, E. A. Toward a theory of task motivation and incentives. *Organizational Behavior and Human Performance*, 1968, **3**, 157–189.

LOCKE, E. A., CARTLEDGE, N., & KNERR, C. S. Studies of the relationship between satisfaction, goal setting, and performance. *Organizational Behavior and Human Performance*, 1970, **5**, 135–158.

LOCKE, E. A., CARTLEDGE, N., & KOEPPEL, J. Motivational effects of knowledge of results: A goal-setting phenomenon? *Psychological Bulletin*, 1968, **70**, 474–485.

RONAN, W. W., LATHAM, G. P., & KINNE, S. B., III. The effects of goal setting and supervision on worker behavior in an industrial situation. *Journal of Applied Psychology*, 1973, **58**, 302–307.

(Received January 8, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 2, 192–196

ASPECTS OF DRIVING PERFORMANCE AS A FUNCTION OF FIELD DEPENDENCE

PAUL L. OLSON [1]

General Motors Research Laboratories, Warren, Michigan

Three studies were conducted to investigate the relationship between field dependence and performance in selected driving situations. The first study compared the field dependent and field independent groups in terms of control in a skidding situation. Overall performance levels were not statistically different although the field independent subjects showed a performance improvement after the first trial and the field dependent subjects did not. The second and third studies compared performance in a platoon car-following situation. Results indicated that field dependent subjects may not use information from vehicles ahead of the lead vehicle as well as do field independent subjects.

This study reports the results of three studies relating field dependence (e.g., Witkin, Dyk, Faterson, Goodenough, & Karp, 1962) to selected measures of driving performance. Although the literature concerning field dependence is substantial, relatively few references relate to the driving situation.

A comprehensive program of research in the application of field dependence to driver capability was proposed by Michael (1968), but no experimental work was reported. Barrett and Thornton (1968) discuss the results of a study in which subjects were exposed to an emergency situation (pedestrian stepping in front of a car) in a driving simulator. They found a measure of field dependence, rod-and-frame test, significantly correlated with reaction time, deceleration rate, hit-miss criterion and incidence of simulator sickness. Field independent subjects performed better in the tests but were more prone to simulator sickness. A later study (Barrett, Thornton, & Cabe, 1969) investigated an alternative measure of field dependence, the Embedded Figures Test, as well as the rod-and-frame test, using the same driving task and replicating their findings with respect to prediction of emergency behavior.

Thornton, Barrett, and Davis (1968) found that field independent subjects performed better in identifying targets in aerial photographs. Similarly, Moses (1970) found that his field independent subjects performed better in a complex vigilance task and Ton (1972) found that field independent subjects responded more quickly in the detection of motion in depth. These results suggest that field independent subjects may also perform better in identifying pertinent stimuli in a complex street scene.

The only published study dealing with field dependence and actual accident involvement is that reported by Harano (1969), who found that field dependent drivers associated with a higher frequency of traffic mishaps than field independent drivers.

The work reported in this study examined the relationship of field dependence and measures of driving performance, specifically, skid control and car following capability.

Subjects

This study was conducted in three phases. For the first two phases a sample was constructed by screening 93 employees of the General Motors (GM) Proving Ground in Milford, Michigan and by using the portable rod-and-frame test described by Oltman (1968). Subjects scoring at the extremes were compared in terms of personal data in an effort to find 10 each field independent subjects and field dependent subjects that were matched as well as possible in terms of age, sex, and driving experience. Rod-and-frame test scores for the 10 field dependent subjects ranged from 50 to 198; those scores for the 10 field independent subjects ranged from 4 to 15.

---

[1] The author wishes to express his gratitude to a number of individuals who assisted at one phase or another of the research reported in this paper. D. N. Goodenough and P. K. Oltman were most helpful at all phases of the project, supplying ideas and making suggestions. R. Herman and R. W. Rothery assisted in the design of the car-following studies and were also quite helpful in other ways. K. Migda assembled and maintained the instrumentation throughout and also assisted in collecting data. Sincere appreciation is also due the two reviewers who made significant contributions to the revision of this paper.

Requests for reprints should be sent to Paul Olson, who is now at the Highway Safety Research Institute, Human Factors Department, University of Michigan, Huron Parkway at Baxter Road, Ann Arbor, Michigan 48105.

For Phase 3 another sample was generated that was drawn from a group of 72 employees at the General Motors Research Laboratories. Similar matching procedures were employed. Scores for the 10 field dependent subjects in this sample ranged from 75 to 179; those scores for the 10 field independent subjects ranged from 3 to 23.

PHASE 1: SKID CONTROL

Control of a car on a slippery surface is one of the most challenging situations the average driver faces, requiring precise control inputs and exact timing. It seemed promising to determine whether performance differences in a skid control task might be related to field dependence.

Method

A straight lane, 7 feet wide and 400 feet long was marked out with traffic cones set at 20-foot intervals on a special low-friction surface at General Motors' Milford Proving Ground. The front brakes on the test vehicle were disabled, making it prone to spin out. Subjects were instructed to approach the lane at 30 mph and begin braking as they came abreast of the first pair of cones, trying to stop in the shortest distance possible while keeping the car within the lane of cones. All 20 of the proving ground subjects described earlier participated in this study. Each made 10 runs through the course. No practice trials were allowed.

Results

Analysis of variance for mean stopping distance indicated that only the interaction between subject classifications and trial was significant ($F = 2.22$, $df = 9$, $p < .05$). An analysis of simple effects revealed significant variance ($F = 2.61$, $df = 9$, $p < .01$) among trials for only the field independent group. A Newman-Keuls test on these data indicated that first trial performance was significantly poorer than all other trials ($p < .01$). An analysis of variance conducted on the cone data revealed no significant differences.

Discussion

Differences between the groups of subjects in performance on this task of skid control could not be established. It is interesting to note the superior ability of the field independent subjects to improve their performance after a brief exposure to such a task. This finding is consistent with the results of a number of studies (see Witkin et al., 1962) that indicate a greater capability on the part of field independent individuals to alter their behavior in the face of changed conditions.

PHASE 2: CAR FOLLOWING

Since many drivers log a significant percentage of their miles in congested traffic, where they are close enough to a lead car to be immediately influenced by what it does, and many drivers maintain headways of less than a second on fast-moving freeways, it would seem essential that they be well informed about what is happening ahead of the car directly in front of them. It has been suggested by two sources (D. R. Goodenough, personal communication, May 1968; Michael, 1968) that the field dependent driver could have his consciousness so dominated by the car immediately in front of him that he may have a reduced capability for detecting significant maneuvers by vehicles further up the line. This was the hypothesis tested in Phases 2 and 3.

Method

Subjects drove the last car of a three-car platoon. Each subject drove under two conditions: (*a*) open—allowing the normal visibility of both lead cars, and (*b*) closed—making an effort to screen the view of the lead car by blocking the rear window of the middle car and by shifting the lateral position of both lead cars to make it more difficult for the subject to see around the middle car.

All 20 of the Proving Ground subjects described earlier participated in this study. All took the conditions in the same order in preference to the usual balanced design, since it was felt probable that

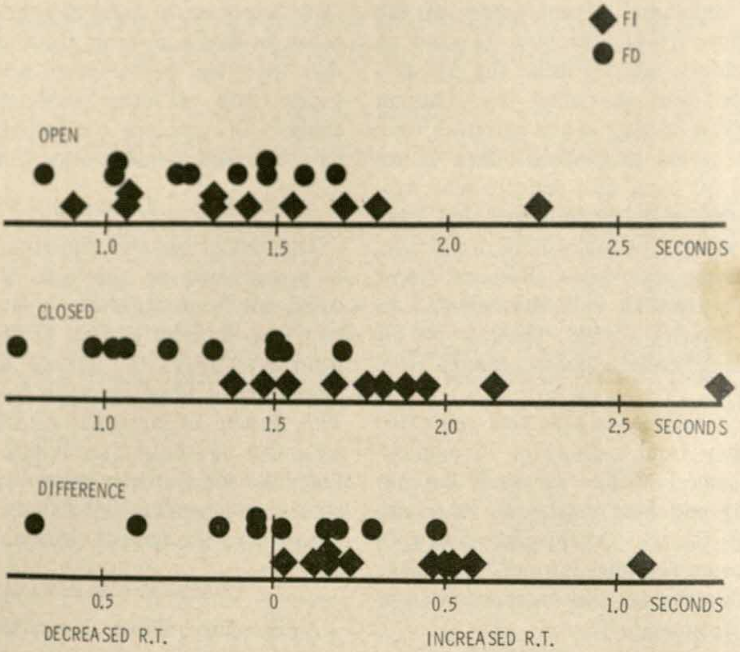


FIGURE 1. Reaction times for individual subjects to onset of accelerations by the lead car.

subjects who took the closed condition first became more attentive than usual in the open condition because the purpose of the study had become apparent.

The following dependent variables were employed:

Reaction time. The driver of the lead car went through a series of programmed mild accelerations and brake applications in a speed range from 20 to 60 mph, and the subject's reaction time to the onset of these maneuvers was measured. It was expected that the more a driver used information gained from cars ahead of the lead vehicle, the more his reaction time should be affected by denying him that information.

Intervehicular separation. The subject's car was connected to the middle car by a "yo-yo" device which provided a continuous analog record of spacing. Reaction time is affected by intervehicular separation and if the two groups differed in this respect, it could have biased the results.

The study was run on a straight road facility 2.6 miles long. Subjects made one complete lap, 5.2 miles, under each condition (open–closed). During each lap they experienced eight accelerations and seven brake applications. The same driving schedule was followed on each lap.

The subjects were read the following instructions:

> Imagine yourself driving home on a crowded urban freeway. The density of traffic is such that you cannot pass or change lanes and you can only follow along behind the vehicle in front of you. In this test we would like you to drive as you think you would under such circumstances. Please bear in mind that in a real situation such as I have described, if you tried to make things easy for yourself by dropping far back, someone would surely cut in front of you. So maintain a reasonable following distance.

Subjects were also informed that their car was connected to the vehicle in front by a fine wire that measured the distance from their car to the car in front. Questions were then answered and the test began.

Results and Discussion

An examination of the data made it clear that despite the precautions taken, about three quarters of the subjects were able to see the lead car at least some of the time during the closed condition. As a result there was no difference in reaction time to brake applications between the groups of subjects. However, reaction time to accelerations did show a significant difference. These data are plotted in Figure 1.

The top part of Figure 1 shows the mean reaction time for each subject under the open condition. The second line shows the mean

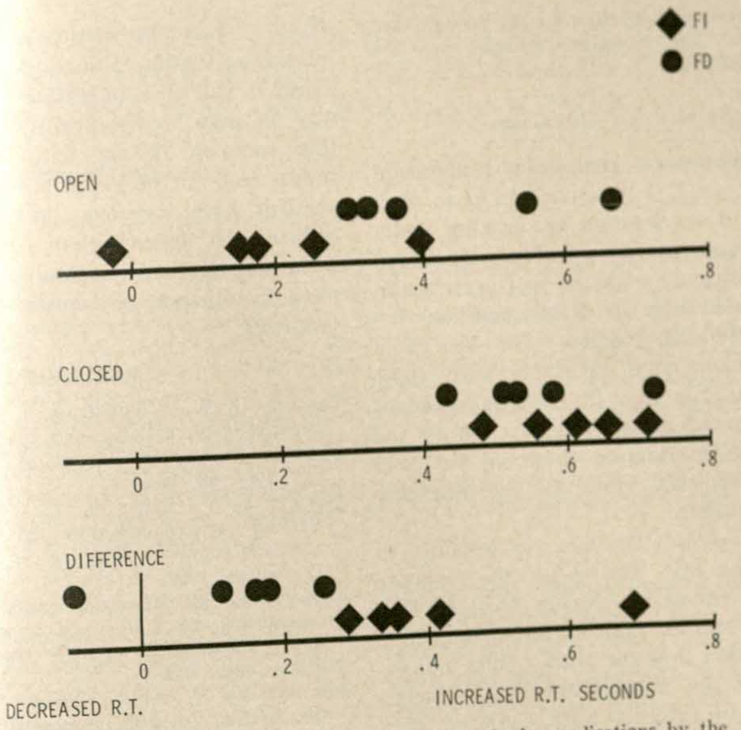


FIGURE 2. Reaction times for individual subjects to brake applications by the lead car.

reaction times associated with the closed condition. Note that the field dependent subjects occupy much the same position on the scale as they did formerly, while the field independent subjects have moved up scale (longer reaction times). The bottom line shows a plot of each subject's reaction time under the closed condition after having subtracted his times under the open condition. Note that half of the field dependent subjects showed increased reaction time under the closed condition, half a decreased reaction time. On the other hand, all of the field independent subjects showed increased reaction times under the closed condition. The separation between the groups was significant ($p < .01$) as determined by the Mann-Whitney U test.

There was a significant difference in mean headways ($p < .01$) between groups as determined by the Mann-Whitney U test, with the field dependent subjects tending to follow more closely than did the field independent subjects.

PHASE 3: CAR FOLLOWING REPLICATION

Because the Phase 2 study permitted incomplete obscuring of the lead car, a replication was conducted to determine whether undesirable artifacts may have resulted.

Method

This study was conducted on the General Motors Technical Center site, using five each field independent and field dependent subjects drawn from the Research Laboratory pool described earlier. The procedure was the same as for Phase 2 with four exceptions: (*a*) greater care was taken in the lateral positioning of the vehicles to more effectively screen the lead car; (*b*) the outboard pair of the four lamps on the rear of the lead car was disabled during the closed trials; (*c*) a four-lane divided facility with .7 mile straightaways was used and the subjects made two complete laps of the track under each condition, always starting in the open condition, next going to the closed condition and finishing with the open condition again; and (*d*) instructions were changed to the extent that the last two sentences ("Please bear in mind. . . .") were deleted.

As in the first study, Phase 1, all measures were taken on the straight portion of the track. There were nine acceleration trials and eight braking trials per condition. The dependent variables were the same as in Phase 2.

Results and Discussion

In a preliminary analysis no significant difference ($p > .05$) was found in mean reaction time to acceleration or braking under the two open conditions. The subsequent analysis dealt only with the first open condition contrasted with the closed condition.

The acceleration reaction time data from this phase compared favorably with those from Phase 2, in that the field independent subjects showed more of a reaction to the loss of lead information than did the field dependent subjects. However, the difference was not significant.

Figure 2 shows the data for response to braking. The first line shows the response under open conditions. Removal of the lead information was associated with a lengthening of the reaction time for both groups, but especially for the field independent subjects, such that all subjects were clustered at a relatively narrow range on the scale. Subtracting one condition from the other reveals that most subjects suffered an increase in reaction time under the closed condition, but the effect on the field independent subjects was sufficiently pronounced in that there was a total separation between the groups.

There was no difference in mean headways between groups in Phase 3.

DISCUSSION

The results of the preceding two studies, Phases 2 and 3, provide some evidence that individuals classified as field dependent are less capable of or less willing to avail themselves of lead information in a platoon car-following situation. Unless the field dependent driver is able to compensate for this in some way, he may be significantly handicapped in this common driving circumstance. These results lead to the expectation that field dependent drivers are over involved in rear-end collisions. It appears that a relatively large-scale investigation of actual collision involvement by different perceptual types could be rewarding.

REFERENCES

BARRETT, G. V., & THORNTON, C. L. Relationship between perceptual style and driver reaction to an emergency situation. *Journal of Applied Psychology*, 1968, **52**, 169–176.

BARRETT, G. V., THORNTON, C. L., & CABE, P. A. Relation between embedded figures test performance and simulator behavior. *Journal of Applied Psychology*, 1969, **53**, 253–254.

HARANO, R. M. The relationship between field dependence and motor vehicle accident involvement. Unpublished master's thesis, Sacramento State College, 1969.

MICHAEL, D. N. Some preliminary hypotheses on the relation of field approach to driving behavior. (I.D.R. No. 437864, Project No. 1066) University of Michigan, Institute for Social Research, January 1968.

MOSES, J. L. Selecting vigilant types: Predicting vigilance performance by means of a field dependence test. *Experimental Publication System*, 1970, No. 4, Ms. No. 151B.

OLTMAN, P. K. A portable rod-and-frame apparatus. *Perceptual and Motor Skills*, 1968, **26**, 503–506.

TON, W. H. Perceptual style and detection of motion in depth. *Perceptual and Motor Skills*, 1972, **34**, 423–428.

THORNTON, C. L., BARRETT, G. V., & DAVIS, J. A. Field dependence and target identification. *Human Factors*, 1968, **10**, 493–496.

WITKIN, H. A., DYK, RUTH B., FATERSON, H. F., GOODENOUGH, D. R., & KARP, S. A. *Psychological differentiation*. New York: Wiley, 1962.

(Received January 12, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 2, 197-201

A VIEW OF WHAT CAN HAPPEN WHEN BEHAVIORAL EXPECTATION SCALES ARE DEVELOPED IN ONE SETTING AND USED IN ANOTHER

WALTER C. BORMAN[1]
Personnel Decisions, Inc., Minneapolis, Minnesota

W. ROBERT VALLON
University of California, Berkeley

Supervisors rated their subordinate nurses using both Smith and Kendall's behavioral expectation scales and a numerically anchored format. A counterbalanced research design was used to explore the differences in the operational effectiveness of the two formats. Results indicated that in terms of interrater reliability and confidence in ratings, the Behavior Expectation Scale was slightly superior. However, when the *simpler* scale was used there was significantly less leniency effect and raters were better able to discriminate among different ratees in terms of performance. It was concluded that when a behavioral expectation scale is transplanted from one setting to another, the effectiveness of the scaled-expectations format may suffer because the raters do not participate in scale development and/or certain anchors are inappropriate for the new situation.

The Behavioral Expectation Scale (BES) format has been praised as a significant improvement over other rating scale formats designed to measure job performance (e.g., Campbell, Dunnette, Lawler, & Weick, 1970; Dunnette, 1966). Although procedures have varied somewhat, Smith and Kendall's (1963) original suggestions for developing behavioral expectation scales are generally followed closely. Briefly, those procedures are as follows: (*a*) A sample of raters contributes behavioral examples representing low, average, and high performance on the job in question. The investigator attempts to exhaust the job-performance domain by sampling the content of the job as completely as possible in terms of these concrete observables. (*b*) The behavioral examples are clustered by content, and dimensions of performance are named and defined. (*c*) Each member of a separate sample of raters scales the performance examples (or anchors) and sorts them into the dimension categories. Means, standard deviations, and dimension frequencies are tabulated for each anchor. (*d*) The investigator decides which anchors are to be included for each dimension based on the criteria of low standard deviations for the scaling task and rater agreement on the sorting task. When two or more anchors have means very near each other and otherwise meet the criteria of membership for a single dimension, the one with the lowest standard deviation is generally chosen. In addition, the retranslation step outlined in (*c*) and (*d*) above provides a further check on the meaningfulness of the performance dimensions chosen in step (*b*) above. (*e*) The investigator retains those dimensions which he feels have a reasonable number of behavioral anchors meeting the criteria stated in (*d*) above.

Since the first Behavioral Expectation Scale was published (Smith & Kendall, 1963, for nurses), a number of other behavioral expectation scales have been developed for a variety of jobs. The purported advantages of the Behavioral Expectation Scale format over other rating scale formats include the following: (*a*) During the evaluation process, behavioral anchors force raters to attend more to each ratee's actual behavior than to a global overall feeling about the ratee. (*b*) Behavioral expectation scales are more meaningful for raters because they uniquely understand the jargon and situations used in the content of behavioral anchors; thus, according to the "meaningfulness hypothesis" (O'Donovan, 1965) behavioral expectation scale ratings should be less susceptible to the central tendency error. (*c*) Behavioral expectation scales provide relatively less ambiguity in rater understanding of each dimension's meaning because of the careful definitions prepared by samples of raters as well as the added definitions supplied by the behavioral anchors on each dimension.

---

[1] Requests for reprints should be sent to Walter C. Borman III, Personnel Decisions, Inc., 2515 Foshay Tower, Minneapolis, Minnesota 55402.

Despite considerable praise for the Behavioral Expectation Scale, to the best of the authors' knowledge, only one study comparing its effectiveness with that of a simpler format has been published (Campbell, Dunnette, Arvey, & Hellervik, 1973). Since the time and effort necessary to construct a behavioral expectation scale is great compared to other formats commonly used, studies which examine the relative strength of behavioral expectation scale formats with simpler formats attempting to measure the same criterion space seem to be in order. Sechrest's (1963) incremental validity notion is relevant here. Before the Behavioral Expectation Scale is declared superior to other formats, investigators should compare behavioral expectation scales with less complex rating methods using dependent variables appropriate to the determination of rating-scale effectiveness.

Instead, *scale reliabilities* have often been submitted as evidence supporting behavioral expectation scales (e.g., Fogli, Hulin, & Blood, 1971; Landy & Guion, 1970; Smith & Kendall, 1963). Scale reliabilities for a given behavioral expectation scale dimension are generally computed by splitting the sample of individuals who scaled the anchors into two equal sized groups. Next, for each group the mean scale values are computed for each anchor on this dimension. Finally, the two sets of mean scale values are correlated (n of the correlation = number of anchors). As others have pointed out (e.g., Borman, 1972; Fogli et al., 1971; Zedeck & Baker, 1972), scale reliabilities are inadequate for predicting a format's operational effectiveness. In fact, this kind of internal reliability can be raised by: (*a*) choosing anchors with small standard deviations; (*b*) choosing anchors whose means are relatively far apart; (*c*) having more raters assign values to the anchors during the retranslation process; and (*d*) including fewer anchors for each dimension. The last three influences, although improving scale reliability, either have nothing to do with scale effectiveness or, in the case of (*d*), actually may lower the potential of the scale to define effectively its criterion space. Even (*a*) above may have nothing to do with operational scale effectiveness. It merely deals with the potential of a scale to be less ambiguous. Thus, scale reliability indices for behavioral expectation scales are not needed nearly as much as data which bear on the operational effectiveness of the scales.

There have been a number of rating-scale-format comparison studies (for a review, see Peters & McCormick, 1966). In general, results suggest that formats which include both dimension definitions and some kind of verbal anchors are superior operationally to formats that do not possess these characteristics. Of particular interest to the authors are the reported findings of Campbell et al. (1973). These investigators compared ratings on a behavioral expectation scale format to those done on a "summated" rating scale format. Store managers and their assistants each rated their subordinate department managers using a newly prepared behavioral expectation scale format and the summated scales. Results of the Campbell et al. (1973) study favored the scaled expectations format. More specifically, the behaviorally anchored scales yielded less leniency error, greater discriminant validity, and a clearer factor-analytic solution than did the summated rating format.

The main purpose of the present study was to investigate the relative effectiveness of a behavioral expectation rating scale applied in a setting different from the one in which it was developed. To accomplish this, the performance of Smith and Kendall's original Behavioral Expectation Scale was compared to the performance of a less complex numerically anchored rating format using five dependent variables seen as important for judging rating scale effectiveness. The dimension names and definitions were the same for both formats, but the simpler scale did not contain the behavioral anchors found on the Behavioral Expectation Scale format.

METHOD

Pairs of raters from 24 subgroups of nurses in two large Northern California hospitals participated in the study. The number of ratees in each subgroup varied from 7 to 29. The median was 16.

Subgroups from each hospital were randomly assigned to one of two categories. Rater pairs from 12 of the groups performed ratings using the Behavioral Expectation Scale (e.g., both raters from the group chosen as number one used the Behavioral Expectation Scale at this point). Raters from the other 12 groups accomplished their performance evaluations using the simpler scale. After approximately three weeks, those who had used the Behavioral Expectation Scale first evaluated the same individuals with the simpler scale. Raters who had been in the simple scale group first switched to the Behavioral Expectation Scale for the second administration. For every rating made, each rater was also asked to record confidence estimates from 1 (not at all confident) to 9 (extremely confident) according to how sure she felt about each evaluation made.

Since interrater reliability can be influenced profoundly by using raters who observe ratees under different conditions (Borman, 1972; Campbell et al., 1970; Zedeck & Baker, 1972), the investigators outlined two requirements for the participation of the rater pairs in the study. First, supervisors at both hospitals were asked to provide rater pairs whose two members possessed similar working relationships with nurse ratees. Second, the investigators requested that the sample contain (for each subgroup) pairs of raters who observed ratees for approximately the same amount of time per day. These two requirements were best satisfied by having head nurses and assistant head nurses from each subgroup participate as the rater pairs. It was felt that for these organizations, satisfying the two stipulations for rater-pair participation discussed previously made interrater reliability a legitimate criterion of scale effectiveness. The other four dependent variables used were means of each rater's evaluations, standard deviations of the evaluations, mean interdimension (halo) correlations for each rater, and mean confidence ratings.

Ten interrater reliability coefficients were available for each participating subgroup, 5 for each pair's first administration on each dimension, and 5 for the second administration on the same dimensions. Each interrater reliability coefficient was transformed to a z score using the Fisher r to z transformation. These z scores were subsequently used in an analysis of variance to be discussed later. Data points for the dependent variable, *means*, consisted of each rater's mean rating level across ratees on a single dimension and administration. Since the scale points differed for the two formats (0–2 for the Behavioral Expectation Scale and 1–9 for the simpler scale), each mean behavioral expectation scale rating for each rater was transformed by multiplying by 4 and adding 1. *Standard deviations'* data points were individual rater's standard deviations of their evaluations across ratees on a single dimension and administration. Standard deviations of each rater's evaluations on the Behavioral Expectation Scale ratings were multiplied by 4 to arrive at indices equivalent to those for the simpler scales. "Halo" data points were established by taking the mean (using Fisher r to z transformations) of each rater's 10 interdimension correlations, first for the BES and then for the simpler scale ratings. Finally, each rater's mean confidence rating for each administration on each dimension was taken as a data point for the confidence level dependent variable.

TABLE 1

BEHAVIOR EXPECTATION SCALE (BES) AND THE SIMPLER SCALE (SS) MEANS ON THE FIVE DEPENDENT VARIABLES

Dependent variables	Behavior expectation scale	Simpler scale
Interrater reliability	.61	.56
Mean level of ratings	6.43[a]	6.14
Standard deviations of ratings	1.02[a]	1.12
Confidence level	6.50	6.38
Dimension intercorrelation (halo)	.73	.73

[a] Means and standard deviations of ratings are adjusted for differences in the number of scale points on the Behavior Expectation Scale and the Simpler Scale.

The research design used for four of the five dependent variables was a three-way (Format × Dimension × Administration) analysis of variance (ANOVA). That is, four separate univariate 5 × 2 × 2 ANOVAs were done with interrater reliability, means, standard deviations, and confidence level as dependent variables. A two-way (Format × Administration) ANOVA was used with the dependent variable, "halo."

RESULTS

Table 1 shows the means on the five dependent variables in the present study for each of the two formats. Table 2 is an analysis of variance (ANOVA) summary for the same five dependent variables.

Although mean interrater reliability was higher for the Behavioral Expectation Scale, the difference was not significant. In addition, mean dimension intercorrelations (halo) were virtually equivalent for the two formats. Differences between format performance in terms of means and standard deviations actually favored the simpler scale. Adjusted mean ratings on the Behavioral Expectation Scale were significantly higher ($p < .01$) than on the simpler scale (Table 2). That is, the simpler scale showed less leniency effect than did the Behavioral Expectation Scale. Stan-

TABLE 2

ANALYSIS OF VARIANCE SUMMARIES FOR THE DEPENDENT VARIABLES: INTERRATER RELIABILITY, MEAN RATING LEVEL, STANDARD DEVIATION OF RATINGS, CONFIDENCE LEVEL, AND INTERDIMENSION CORRELATIONS

Sources of variance	*df*	Interrater reliability		Mean level of ratings		Standard deviations of ratings		Confidence level		Interdimension correlations (halo)[a]	
		MS	*F*	*MS*	*F*	*MS*	*F*	*MS*	*F*	*MS*	*F*
Formats (A)	1	.29		9.43	8.95**	1.10	5.00*	1.50		.00	
Dimensions (B)	4	.37		3.31	3.14*	.19		1.20		—	
Administrations (C)	1	.49		.15		.27		8.66	7.20**	.06	
A × B	4	.71		.19		.04		.03		—	
A × C	1	.48		8.59	8.15**	.23		1.66		.19	
B × C	4	.40		.11		.02		.30		—	
A × B × C	4	.55		.42		.09		.17		—	

Note. The first author will supply further details of the ANOVA on request. Only significant *F*s are given.
[a] The ANOVA with "halo" as the dependent variable was a two-way, 2 × 2 (Formats × Administrations) design.
* $p < .05$.
** $p < .01$.

dard deviations of ratings on the simpler scale were significantly ($p < .05$) greater than the adjusted standard deviations of evaluations done on the Behavioral Expectation Scale. Thus, raters tended to spread out their ratings more on the simpler scale. Although this is not necessarily a virtue, scales that encourage raters to fairly discriminate among the performance of different ratees to a greater extent than do other scales are generally considered superior. Finally, raters were more confident in their evaluations using the Behavioral Expectation Scale than they were using the simpler scale, but not significantly so.

DISCUSSION

In a situation where interrater agreement seemed to be a fair benchmark for judging rating-scale effectiveness, no significant difference was found between the Behavioral Expectation Scale and a much simpler format on this criterion. In addition, because of the high mean "halo" for both formats, it appeared that most ratings were a reaction to overall ratee favorableness without any careful distinction made of performance on different dimensions. Although behavioral expectation scales should be especially good at lowering "halo" because of the relatively specific behavior orientation of the scales, dimension intercorrelations were approximately the same for both formats. In short, under these conditions, the Behavioral Expectation Scale exhibited no superior operational characteristics.

One possible explanation for the lack of superiority shown by the Behavioral Expectation Scale in this study is that nurse raters may have had trouble relating to certain out-of-date behavioral examples contained within the behavioral expectation scale used in this study. A supervisor nurse did suggest to the second investigator that a few of the anchors were no longer appropriate as performance examples for nurses in 1972. In addition, raters did not have the opportunity to prepare their own dimensions and anchors. Rater participation in scale preparation may be an important variable for determining rating format effectiveness. Greater participation in the performance-evaluation process may increase the raters' motivation to evaluate individuals carefully. As Campbell et al. (1973) have indicated, development of a behavioral expectation scale instrument seems to be a valuable learning experience for the raters. Participation forces the raters to define carefully effective performance and assess ratees accordingly.

Perhaps the clearest message communicated by the results of this investigation is that a practitioner who wishes to use a behavioral expectation scale which has been developed in a different setting cannot assume that the scales give him any better criterion information than does a much simpler format. That is, the psychometric superiority of the Behavioral Expectation Scale method under conditions in which the raters themselves develop the rating scales has not been directly challenged by the results in this study. Instead, the findings suggest that more research is needed to determine the conditions under which a behavior expectation rating system might be psychometrically superior to simpler performance-evaluation formats.

REFERENCES

Borman, W. C. The rating of individuals in organizations: An alternative approach. Unpublished doctoral dissertation, University of California, Berkeley, 1972.

Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L. W. The development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology*, 1973, **57**, 15–22.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., III, & Weick, K. E., Jr. *Managerial behavior, performance, and effectiveness*. New York: McGraw-Hill, 1970.

Dunnette, M. D. *Personnel selection and placement*. Belmont, California: Wadsworth, 1966.

Folgi, L., Hulin, C., & Blood, M. Development of first level behavioral job criteria. *Journal of Applied Psychology*, 1971, **55**, 3–8.

Landy, F. J., & Guion, R. M. Development of scales for the measurement of work motivation. *Organizational Behavior and Human Performance*, 1970, **5**, 93–103.

O'Donovan, D. Rating extremity: Pathology or meaningfulness? *Psychological Review*, 1965, **72**, 358–372.

Peters, D. L., & McCormick, E. J. Comparative reliability of numerically anchored versus job-task anchored rating scales. *Journal of Applied Psychology*, 1966, **50**, 92–96.

Sechrest, L. Incremental validity: A recommendation. *Educational and Psychological Measurement*, 1963, **23**, 153–158.

Smith, P. C., & Kendall, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, 1963, **47**, 149–155.

Zedeck, S., & Baker, H. T. Nursing performance as measured by behavioral expectation scales: A multitrait-multirater analysis. *Organizational Behavior and Human Performance*, 1972, **7**, 457–466.

(Received January 15, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 2, 202–206

WHAT DOES THE LEAST PREFERRED CO-WORKER SCALE REALLY MEASURE?:

A COGNITIVE INTERPRETATION

MARTIN G. EVANS[1]
London Business School, England

JERRY DERMER
Sloan School of Management, Massachusetts Institute of Technology

The Least Preferred Co-Worker Scale has been variously interpreted as a measure of leadership style or as a measure of the cognitive complexity of the leader. In this study, using 112 subjects including business students, managers and systems analysts, the Least Preferred Co-Worker scale scores were correlated with two other measures of differentiation and with three cognitive measures in an attempt to assess the interpretation of the Least Preferred Co-Worker Scale as a cognitive measure. Results show that while the low end of the Least Preferred Co-Worker Scale seems to be associated with cognitive simplicity, the high-least preferred co-worker subject was not unequivocably cognitively complex.

Despite extensive research over the past 20 years, interpretation of Fiedler's (1958, 1967) Least Preferred Co-Worker Scale (LPC) remains elusive. Some have suggested that a manager's LPC score can be interpreted as a direct indicator of his leadership style (Fiedler, 1967, 1971). Others, relying on the fact that LPC scores have been found to correlate with several measures of cognitive complexity (Steiner & McDiarmid, 1957; Vannoy, 1965; Weissenberg & Gruenfeld, 1966), have advanced cognitive-based leadership theories explaining the relationship of LPC scores to leader behavior (Foa, Mitchell, & Fiedler, 1971; Mitchell, 1970). Yet, despite these recent attempts to raise LPC-based leadership studies to a higher theoretical plane, there is still considerable doubt surrounding the cognitive interpretation of the LPC scale.

Interpreting LPC scores as measures of the ability to discriminate links them to cognitive complexity theory. The scale consists of 16 items presented in semantic differential format. Respondents are asked to describe their least preferred co-worker by indicating where he should be positioned along an eight-interval continuum anchored by the polar adjectives. These responses measure a respondent's ability to differentiate among the characteristics of an individual as a person (e.g., items such as cold–warm; friendly–unfriendly) and also as a co-worker (e.g., items such as enthusiastic–unenthusiastic; efficient–inefficient). Since the ability to make fine discriminations characterizes cognitively complex individuals (Schroder, Driver, & Streufert, 1967) the LPC scale has been interpreted to be a measure of cognitive complexity.

A problem arises with this interpretation, however, when one attempts to define the nature of the differentiation process captured by the LPC scale. Initially, differentiation was measured by an assumed similarity between opposites measure, computed as the square of the difference between the mean rating of one's most preferred co-worker and one's least preferred co-worker. Later, it was found that scoring on the most preferred co-worker scale was consistently high and that the assumed similarity between opposites measure and LPC scores correlated consistently; therefore, only the LPC score was retained (Fiedler, 1967). Thus, the average score on the scale items describing a least preferred co-worker became the LPC score. Low scores, resulting when a respondent stereotyped a poor co-worker as also being an undesirable person, were interpreted as indicating a tendency to dichotomize people into good or bad categories, a tendency attributed to conceptually simple individuals.

---

[1] The authors wish to acknowledge the helpful comments of U. G. Foa, D. T. Hall, R. Mansfield, T. R. Mitchell, R. I. Simon, and H. C. Triandis on an earlier draft of this paper.

Requests for reprints should be sent to Martin G. Evans, who is now at the Faculty of Management Studies, University of Toronto, Toronto M5S IV4, Canada.

Steiner and McDiarmid (1957) demonstrated that Fiedler's assumed similarity between opposites score was jointly determined by an elevation (mean score) effect and a profile dissimilarity effect, the latter measured as the square of the difference between each most preferred co-worker- and least preferred co-worker-item score summed over all the items of the scale. High scores on their profile-dissimilarity measure were interpreted as reflecting a high degree of differentiation. Furthermore, based on the finding that elevation and profile-dissimilarity effects were independent of each other, it was suggested that their two measures tapped different types or styles of differentiation.

Recently, Mitchell (1970) and Foa et al. (1971) have suggested that the 16-item LPC scale (Fiedler, 1967) can be partitioned into three distinct subscales: (*a*) tapping a task (3 items), (*b*) an interpersonal (11 items) as well as a task/interpersonal, or (*c*) mixed dimension (2 items). The mean scores on these three dimensions were used to investigate the ability to discriminate which was measured in terms of dimensional differentiation. That is, the relationship among the scores on these dimensions was interpreted to be the indicator of the extent to which an individual stereotypes or differentiates his least preferred co-worker.

Thus, data from the LPC instrument have been used to produce three different indexes of differentiation: (*a*) elevation, (*b*) dimensional differentiation, and (*c*) profile dissimilarity. Each of these indexes reflects a different type or degree of differentiation and hence each can be considered to be a correlate of cognitive complexity. But leadership studies have generally relied only on LPC (i.e., mean or elevation) scores as the measure of the ability to differentiate.

This study is designed to contribute to the understanding of the LPC score as a measure of differentiation and hence as a correlate of cognitive complexity by determining (*a*) the extent to which differentiation measured by LPC is also related to dimensional discrimination and profile dissimilarity and (*b*) the extent to which LPC relates to other measures of cognitive complexity.

METHOD

Sample

Subjects included 112 night school students in business (both master's of business administration (MBA) graduates and undergraduates), managers, and systems analysts. About 50% of the sample had a university degree, and the median age was 30–35.[2]

Instruments

The most recent version of the LPC scale (Fiedler, 1967) was used. Respondents also completed three other scales designed to measure cognitive complexity. The short form (Trodahl & Powell, 1965) of the scale developed by Rokeach (1960) was used to measure authoritarianism and closed-mindedness (dogmatism). To measure intolerance of ambiguity, the scale developed by Budner (1962) was used, and Brim's (1955) scale was used to measure desire for certainty. Individuals who are cognitively complex are expected to be nondogmatic and tolerant of ambiguity (low scores) and comfortable with uncertainty (high scores).

RESULTS

LPC as a Measure of Differentiation

The relationship of the LPC scale to dimensional discrimination was investigated by splitting the sample at the mean LPC score and correlating mean scores on the task, interpersonal and mixed dimensions with the LPC scores for the high and the low LPC subsamples. It was found that for the low LPC subsample, the task and mixed dimensions correlated significantly ($p < .001$, $n = 62$), as did the mixed and interpersonal dimensions ($p < .05$). In the high LPC subsample, however, no interdimensional correlation was significant at the .05 level ($n = 50$). On the whole, these results support the position of Foa et al. (1971) that the high-LPC individual discriminates by using different cognitive dimensions. Thus, it appears that LPC scale does measure dimensional discrimination. It should be noted, however, that this finding only holds for Fiedler's (1967) version of the scale. Work with an earlier version of the scale showed that moderate least preferred co-worker subjects (as measured by this early scale) were the high differentiators (Evans, 1973).

---

[2] As age and education are often related to cognitive dimensions (see Ondrack's (1971) work on dogmatism). For all the analyses, partial correlations were computed controlling for these variables. This did not affect the pattern of interrelationships reported in this article.

To measure profile discrimination, a standard deviation was computed for each respondent's scores on the 11 interpersonal items of the LPC scale. The correlation of this measure with LPC scores for the whole sample was .29 ($p < .01$, $n = 112$), while the correlations with the high and low LPC samples were $-.14$ (*ns*, $n = 50$) and .31 ($p < .05$, $n = 62$), respectively. Thus, although the significant overall correlation indicates that LPC measures profile differentiation, the algebraic signs of the subsamples indicate that the relationship is nonlinear.

However, the correlations between LPC scores and the above-mentioned type of discrimination measure may be artifactual. Measures of profile discrimination based on the sum of squared absolute scores or of squared deviations from the mean may be a subtle function of the LPC scale (i.e., mean score) because the extent of profile discrimination possible is constrained by the mean score. For example, respondents with means in the 4–5 range (on an 8-point scale) can employ ratings of 1 or 8 as long as they are evenly distributed, while those with very high or very low scores cannot use as great a range of discrimination and still produce the same LPC score. Thus, the maximum possible profile variation is a symmetrical curvilinear function of LPC scale with a maximum score of 4.5. Hence, this may account for the curvilinear relationship found.

To correct this bias, the measure of profile discrimination was converted from an absolute to a relative measure by dividing each standard deviation by the maximum possible standard deviation that could be achieved for that LPC score. Correlations were then computed between this new measure and LPC score, both for the sample as a whole as well as the high and the low subsamples. These correlations were in the range of $-.6$ to .3 and were not significant. Thus, it appears that when measures of profile discrimination are corrected for attenuation in the range of possible variation, they are independent of LPC scores. Hence, LPC and profile discrimination measures appear to tap two different types of discrimination.

Based on these results, it appears that while the LPC scale may reflect dimensional discrimination, it is not related to profile differentiation. Thus, it is not a robust indicator of the ability to discriminate one's least preferred co-worker.

LPC Scale as a Correlate of Cognitive Complexity

Intercorrelations between the LPC scores and the cognitive measures revealed several interesting relationships. It was found that high dogmatics tended to score low on the LPC scale ($r = .21$, $p < .05$, $n = 112$), while neither the intolerance for ambiguity nor the tolerance for certainty measures correlated significantly ($r = -.04$ and .02, respectively). Among the cognitive measures themselves, dogmatism and ambiguity correlated significantly ($p < .001$) while the uncertainty measure did not correlate significantly with either.

Although the uncertainty measure was unrelated to the LPC scores, it did appear to moderate the LPC–dogmatism relationship slightly. The pattern of means (low dogmatism–high tolerance for uncertainty = 3.84; low dogmatism–low tolerance for uncertainty = 3.88; high dogmatism–high tolerance for uncertainty = 3.85; high dogmatism–low tolerance for uncertainty = 3.48—the latter cell showed a trend at the $p = .10$ level of being different from the other three cell means) suggests that the high least preferred co-worker person is high in tolerance for uncertain situations, coupled with high or low dogmatism; in addition, he may be low in dogmatism coupled with low tolerance for uncertainty. Thus, the high least preferred co-worker person can be either cognitively complex (low dogmatism, high tolerance for uncertainty) or only moderately so. On the other hand, the low least preferred co-worker individual tends to be characterized by cognitive simplicity–high dogmatism and an intolerance of uncertainty. Thus, it appears that the LPC scale is not a strong correlate of cognitive complexity. In particular, it appears that high and low scores on the LPC scale may not differ in degree but may relate to different cognitive attributes—the low end definitely being a measure of low complexity and stereotyping, the high end being somewhat undefinable.

To explore these possibilities further the sample was again split at the mean LPC score and the LPC scores for each subsample were correlated with the three cognitive variables. These results indicated that the significant overall negative correlation of LPC scores with dogmatism is contributed solely by a significant negative correlation between these variables ($p < .05$) for the low LPC subsample. The high LPC subsample correlation was not significant. Thus, it appears that very low least preferred co-worker subjects are clearly dogmatic and that high least preferred co-worker subjects may either be low or high dogmatics.

Correlations of the LPC scores with ambiguity and uncertainty computed for high and low LPC subsamples were not significant and hence did not clarify the relationship of LPC scores to these measures. On the whole, therefore, this evidence suggests that the low least preferred co-worker respondent lacks complexity and the ability to differentiate. However, the high least preferred co-worker subject may vary along the complexity dimensions.

Discussion

These findings do not provide favorable support to the interpretation of the LPC scale as a measure of cognitive complexity. The results suggest that the low least preferred co-worker subject is a poor differentiator and that the high least preferred co-worker subject may be poor or good at differentiating. The implications of this finding may shed some light on the criticism by Graen, Orris, and Martella (1970) that the correlations in the octants of Fiedler's contingency model (Fiedler, 1967) often do not cluster appreciably. While Graen et al. (1970) and Mitchell et al. (1970) suggest that this lack of clustering may reflect the lack of a good measurement of the situational characteristics (leader power, leader–member relationships, and task complexity), the recent study by Graen, Orris, and Alvares (1971) seems sufficiently well controlled to ensure that cells on the continuum were appropriately filled, and again discrepancies between the correlations in each cell were found.

The research reported in this study suggests an alternative explanation for these differences. It has been found that a low LPC score was consistently an indicator of cognitive simplicity in that it tended to be associated with the combination of high dogmatism and high intolerance for uncertainty. However, the high least preferred co-worker individual could be one of several types: (*a*) cognitively complex—undogmatic and comfortable with uncertainty; and (*b*) cognitively mixed—undogmatic but uncomfortable with uncertainty, and dogmatic but comfortable with uncertainty. The predictions made by Foa et al. (1971) about the strengths of the correlations (between LPC scores and performance) in the various octants therefore only holds when the high least preferred co-worker individual is cognitively complex—undogmatic and tolerant of uncertainty. The presence (in any experiment, in any octant) of high least preferred co-worker leaders of lower complexity (cognitively mixed) weakens the predicted relationship (makes a predicted negative relationship less negative or makes a predicted positive relationship less positive) and might, therefore, explain the range of correlations found by Fiedler (1967) in each octant.

REFERENCES

Brim, O. G. Attitude content-intensity and probability expectations. *American Sociological Review,* 1955, **20**, 68–76.

Budner, S. Intolerance of ambiguity as a personality variable. *Journal of Personality,* 1962, **30**, 29–50.

Evans, M. G. A leader's ability to differentiate, the subordinate's perception of the leader, and the subordinate's performance. *Personnel Psychology,* 1973, **26**, 385–395.

Fiedler, F. E. *Leader attitudes and group effectiveness.* Urbana, Ill.: University of Illinois Press, 1958.

Fiedler, F. E. *A theory of leadership effectiveness.* New York: McGraw-Hill, 1967.

FIEDLER, F. E. Personality and situational determinants of leader behavior. Paper presented at the symposium "Contemporary developments in the study of leadership." Southern Illinois University, 1971.

FOA, U. G., MITCHELL, T. R., & FIEDLER, F. E. Differentiation matching. *Behavioral Science*, 1971, **16**, 130–142.

GRAEN, G., ALVARES, K., ORRIS, J. B., & MARTELLA, J. A. Contingency model of leadership effectiveness: Antecedent and evidential results. *Psychological Bulletin*, 1970, **74**, 285–296.

GRAEN, G., ORRIS, J. B., & ALVARES, K. M. Contingency model of leadership effectiveness: Some experimental results. *Journal of Applied Psychology*, 1971, **55**, 196–201.

MITCHELL, T. R. Leader complexity and leadership style. *Journal of Personality and Social Psychology*, 1970, **16**, 166–174.

MITCHELL, T. R., BIGLAN, A., ONCKEN, G. R., & FIEDLER, F. E. The contingency model: Criticism and suggestions. *Academy of Management Journal*, 1970, **13**, 253–268.

ONDRACK, D. A. Attitudes toward authority. *Personnel Administration*, 1971, **34**(3), [illegible]7.

ROKEACH, M. *The open and closed mind*. New York: Basic Books, 1960.

SCHRODER, H. M., DRIVER, M. J., & STREUFERT, S. *Human information processing*. New York: Holt, Rinehart & Winston, 1967.

STEINER, I. D., & MCDIARMID, C. G. Two kinds of assumed similarities between opposites. *Journal of Abnormal and Social Psychology*, 1957, **55**, 140–142.

TRODAHL, V. C., & POWELL, F. A. A short form dogmatism scale for use in field studies. *Social Forces*, 1965, **44**, 211–214.

VANNOY, J. S. Generality of cognitive complexity-simplicity as a personality construct. *Journal of Personality and Social Psychology*, 1965, **2**, 385–396.

WEISSENBERG, P., & GRUENFELD, L. W. Relationships among leadership dimensions and cognitive style. *Journal of Applied Psychology*, 1966, **50**, 392–395.

(Received January 15, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 2, 207–212

PREDICTING MANAGERIAL SUCCESS OF MASTER OF BUSINESS ADMINISTRATION (MBA) GRADUATES[1]

ALAN G. WEINSTEIN[2]
Carnegie-Mellon University

V. SRINIVASAN
University of Rochester

Annual compensation of 136 alumni of a graduate management program was predicted by using years of work experience, graduate grades and data obtained prior to graduate school admission. Stepwise multiple regression analyses revealed that work experience was more predictive of staff than line compensation. After adjusting for the effect of work experience, graduate grades were significantly correlated to both line and staff earnings though the incremental predictive validity was much higher for line managers. Several pregraduate academic and nonacademic variables were also significant predictors. The results support the ability to predict managerial compensation prior to career and graduate school entry.

Many believe that graduate education in management is instrumental in upgrading the skills of our managerial population. Perhaps the strongest evidence of this is the high demand for master of business administration (MBA) graduates by leading business firms. Such a belief, however, has recently been challenged in a study by Livingston (1971) which asserted that formal managerial education did little to prepare students for successful managerial careers since grades in neither undergraduate nor graduate school correlate with success in management.

This critique deserves analysis. First of all, correlation analysis determines only the *differential* effect of academic performance on business success. If graduate training were to have a more or less *uniform* effect on all graduates, then the correlation is nearly zero and it is erroneous to conclude that graduate training has no effect on managerial performance. Second, part of the evidence used by Livingston came from a comparison of senior executives attending an advanced management program at Harvard with Harvard MBA graduates having comparable years of work experience. The former group earned about a third more, on the average, than the latter which, according to Livingston, was sufficient evidence to question the value of the MBA program. Comparing the earnings of the high performers of one group with those of the mean of another did not deter Livingston from making his point. To make a valid assessment of the treatment effect of MBA training, a comparison of a sample of MBA graduates with a representative sample of non-MBA graduates with similar backgrounds is needed.

The authors could find no published research on the treatment effect of MBA training on subsequent managerial success. In addition, there has been relatively little published research focusing on the differential relationship of formal managerial education with career success. Significant relationships have been reported between managerial compensation and grades in elective MBA courses (Marshall, 1964; Williams & Harrell, 1964) as well as second year MBA grades (Harrell, 1972). In addition, Campbell, Dunnette, Lawler, and Weick (1970) cite research in which undergraduate academic performance related to several criteria of managerial success.

The purpose of the present study is to (*a*) examine the differential effect of MBA performance on career success, and (*b*) develop an early identification model to aid in the selection of students for an MBA program. Specifically, the relationship was explored between undergraduate performance, extracurricular activities and other nonacademic variables, and subsequent managerial success. To evaluate the differential effect of graduate school training, business school graduate grade point average (GPA) was used in conjunction with the preadmission variables in predicting managerial compensation.

---

[1] An abstract of this study was presented at the 79th Annual Convention of the American Psychological Association, Washington, D.C., 1971.

[2] Requests for reprints should be sent to Alan G. Weinstein, Graduate School of Industrial Administration, Carnegie-Mellon University, Schenley Park, Pittsburgh, Pennsylvania 15213.

METHOD

Criterion

From an alumni survey conducted at the Graduate School of Industrial Administration, of Carnegie-Mellon University in 1969, several criteria of managerial success were obtained such as current annual compensation and organizational level attained. Srinivasan, Shocker, and Weinstein (1973) had three groups of senior executives compare profiles of managers on these criteria and make overall judgments of relative success. A composite criterion that was developed, using these judgments, revealed that current compensation was regarded as the single most important criterion. Laurent (1970) also arrived at the same conclusion after analyzing several criterion variables by factor analysis. In addition to this empirical support, annual compensation has "face validity" as a criterion of success to managers and to faculty in business schools. Therefore, annual compensation was used as the measure of managerial success in this study.

Sample

Alumni graduating between 1960–1966 were chosen for the study. The rationale for eliminating 1967–1969 alumni from the sample was based on the desire to reduce spurious effects of grades on current compensation. Graduate grades are correlated with initial earnings perhaps because faculty use these grades to guide their recommendations, and companies use these recommendations and transcripts to guide their selection and compensation. After a few years of experience, however, this effect should diminish and performance on the job would be expected to have a greater effect on current compensation. For each year of the sample investigated, correlational analyses revealed no significant relationships between initial and current compensation.[3]

The criteria for inclusion in the sample were: (*a*) employment in a managerial position, (*b*) employment in the United States, and (*c*) existence of data on preadmission variables, graduate grades and managerial compensation. A preliminary analysis showed differential compensation histories for line and staff managers. In addition, they differed in the nature of the tasks they performed, thus suggesting separate analyses.

Line and staff designations were obtained by having each questionnaire respondent determine to which classification he belonged. The broad definitions given were: (*a*) staff—"No formal managerial responsibility." Typical job titles were analysts, accountants, consultants, and specialists. (*b*) line—"Formal managerial responsibility." Job titles included various types of supervisors, directors, managers, and company officers. The sample included 136 persons; 74 in staff and 62 in line jobs.

Predictors

The predictor variables used in this study were as follows:

Predictor Variables

Administration Test for Graduate Study in Business (ATGSB) scores (Verbal, Quantitative, Total)
Candidate Excellence by School[4] indexes (Verbal, Quantitative, Total)
Undergraduate grade point average by year (freshman, sophomore, junior)[5]
Undergraduate grade point average by subject area (mathematics and statistics, science, engineering, economics, humanities, and social sciences excluding economics, management, computer science)
Undergraduate college major[6] (same areas as above, excluding computer science)
Rating of scholastic recognition
Rating of offices held
Rating of involvement in social organizations
Rating of involvement in sports activities
Rating of letters of recommendation
Weighted graduate study (by length and relevance to management)
Age
Marital status (see Footnote 6)
Graduate (MBA) grade point average

Most of these variables were taken directly from the transcripts and applications. However, relevant information about nonacademic variables existed in either narrative or descriptive form so that measures of these descriptions needed to be developed.

Extracurricular activities. Four independent scales were developed: (*a*) a 9-point scale of scholastic recognition; this included awarded scholarships, membership in honorary societies, Dean's List, etc.;

---

[3] Other studies by Brenner and Lockwood (1965) and Haire, Ghiselli, and Gordon (1967) have found a moderate to high correlation between initial and later compensation for the span of years included in this study. In contrast to the present investigation, both of these studies analyzed salary relationships within single firms, and the high correlations found may partly be due to the effect of a common compensation system.

[4] This index was taken from tables compiled by the Educational Testing Service on the average ATGSB score for all students taking the test from the candidate's undergraduate school.

[5] Undergraduate grade point average (senior) was not available in most cases, since admission decisions were made prior to completion of the senior year.

[6] These predictors were treated as dummy variables.

(*b*) a 9-point scale of offices held in organizations; a higher rating was given to those who held leadership positions in a broad range of organizations; (*c*) a 9-point scale based on the extent of involvement and breadth of social-organizational activities; and (*d*) a 3-point scale of involvement in sports activities ranging from casual playing of athletic games to college level participation.

Letters of recommendation. This rating, on a 9-point scale, was based on the strength and consistency of recommendations as well as explicit mention of certain qualities such as leadership potential and motivation. Two graduate students trained to identify these qualities rated a sample of 30 cases independently. The interrater reliability for these cases was $r = .90$.

Grades. All undergraduate grade point averages (GPA) were computed on a 4-point scale with $A = 4$, $B = 3$, $C = 2$, $D = 1$, and $\text{Failure} = 0$ for each course. Graduate (MBA) grade point average was computed on a 9-point scale with $A+ = 9$, and $D = 0$. To assign courses to content areas, emphasis was placed on the skills used and the nature of the content covered in a course. Course descriptions were used to guide judgments whenever possible.

Adjusting Compensation for Effect of Work Experience

Since current annual compensation was used as the criterion, the number of years of work experience was expected to have an important effect, thus necessitating some form of correction.

For staff managers a plot of current compensation against the number of years since graduation revealed a linear relationship ($r = .58$). For these managers, the years of work experience before coming to graduate school was also significantly related to current compensation. A multiple regression on current annual compensation (Y) with years of work experience before graduation (WEBG) and after graduation (WEAG) as independent variables gave the following result:

$$Y = 10{,}990 + (1{,}120 \times \text{WEBG}) + (873 \times \text{WEAG}).$$

Consequently, adjusted current annual compensation (Y_{adj}) can be written as:

$$Y_{adj} = Y - (1{,}120 \times \text{WEBG}) - (873 \times \text{WEAG}).$$

In the case of line managers, the plot of current compensation against the number of years since graduation approximated a linear relationship up to an experience level of 6 years, but this relationship tailed off for alumni with seven to nine years of experience. Studying this effect more exactly would have involved a multiple regression with seven (dummy) variables, each corresponding to a different level of experience (i.e., three, four, . . . , nine years).[7] Since some of the graduating classes had too few alumni, they were placed into four groups corresponding to (*a*) eight to nine years, (*b*) six to seven years, (*c*) four to five years, and (*d*) three years of experience. A multiple regression with the current compensation (Y) as the dependent variable and these four dummy variables (D_1, D_2, D_3, D_4) as independent variables gave the following result.[8]

$$Y = 20{,}511 + 1{,}873D_1 + 1{,}838D_2 - 1{,}041D_3 - 2{,}670D_4.$$

An F test indicated that the above equation accounted for significantly more variance ($p < .05$) than a linear regression on the same data. In the case of line managers, it was not necessary to adjust for years of work experience before graduation, since there was no significant effect. Consequently, the adjusted compensation (Y_{adj}) for line managers was computed as:

$$Y_{adj} = Y - 1{,}873D_1 - 1{,}838D_2 + 1{,}041D_3 + 2{,}670D_4.$$

RESULTS

The results of (double) cross-validated stepwise multiple regression analyses are reported in Tables 1 and 2 for staff and line managers, respectively. The columns designated as (1), (2), and (3) summarize the results of the following stepwise regressions: (1) predicting unadjusted compensation using preadmission, graduate grade point average and years of experience variables, (2) predicting adjusted compensation using preadmission variables and graduate grade point average, and (3) predicting adjusted compensation using preadmission variables alone. Column (3) represents an early identification model of career success at the time of entry into graduate school. The difference in the explained variance between columns (2) and (3) represents the incremental validity of graduate school performance.

The number of years of work experience had a substantially greater effect on staff managers than line managers. This is demonstrated by the differences in standard deviations of adjusted and unadjusted compensation for staff (2,683 versus 3,722) and line (5,412 versus 5,707). Also, as demonstrated in Table 2, a substantial improvement in pre-

---

[7] See Cohen (1968) for a discussion of dummy variable regression models. A dummy variable D is a binary variable indicating presence ($D = 1$) or lack of presence ($D = 0$) of an attribute. This technique allows for estimating nonlinear effects of a variable.

[8] For a manager with six years experience, $D_1 = 0$, $D_2 = 1$, $D_3 = 0$, and $D_4 = 0$; for a manager with three years experience, $D_1 = 0$, $D_2 = 0$, $D_3 = 0$, and $D_4 = 1$.

TABLE 1

REGRESSION RESULTS ON (STAFF) MANAGERIAL COMPENSATION

Variables	$\bar{X}$	SD	(1)	(2)	(3)	Zero-order correlation with adjusted compensation (on entire sample)
			Beta Weights (on entire sample)			
			With 1969 compensation as dependent variable	With 1969 compensation adjusted for years or work experience	Same as (2) but without using graduate GPA as predictor	
Years of work experience after MBA	5.37	1.95	.48***	—	—	—
Years of work experience prior to MBA	1.51	1.40	.39***	—	—	—
Graduate GPA	6.27	.73	.13	.18	—	.24**
Ratings of involvement in social organizations	4.85	2.03	.14	.19	.24**	.22*
Engineering major	.54	.50	.17*	.23*	.23*	.15
Mathematics and statistics major	.10	.30	.15*	.21*	.22*	.12
Current compensation ($)	17,363[a]	3,722[a]	—	—	—	—
df	—	—	67	69	70	72
R[b]	—	—	.75***	.38***	.34***	—
(Double) Cross-Validated R[c]	—	—	.63***	.31***	.26**	—

Note. Columns designated as (1), (2), and (3) summarize the stepwise regressions. MBA = master of business administration; GPA = grade point average.

[a] After adjusting for the number of years of work experience, the mean and standard deviation were 10,989 and 2,683, respectively.

[b] Adjusted for degrees of freedom, using a standard correction formula (Nunnally, 1967, p. 164).

[c] The total sample was divided by placing odd numbered managers into subsample 1 and even numbered managers into subsample 2. Multiple regression models 1 and 2 were developed from the corresponding subsamples. Predicted criterion scores were obtained for subsample 1 using Model 2 and for subsample 2 using Model 1. The double cross-validated R is the correlation between predicted and actual criterion scores for the entire sample and is a conservative estimate compared to other cross-validation procedures.

* $p < .10$.
** $p < .05$.
*** $p < .01$.

dictability of staff compensation occurs when work experience is included as a predictor. For line managers, there is only marginal improvement.

In the population studied, there were clear findings to support the ability to identify successful managers prior to career entry. Staff managers differed considerably from line managers in the prediction of success. Graduate grades were significantly correlated with adjusted compensation for both line ($r = .49$, $p < .01$) and staff ($r = .24$, $p < .05$) managers, though the incremental predictive validity was much higher for line managers. For staff managers, involvement in social organizational activities was predictive ($p < .05$) whereas for line managers, involvement in sports activities was predictive ($p < .05$). Several other variables predicted adjusted compensation: junior grades (line) ($p < .01$), mathematics and statistics major (line) ($p < .05$) and (staff) ($p < .10$), science grade point average (negatively for line) ($p < .05$), and engineering major (staff) ($p < .10$).

DISCUSSION

The significant correlation of graduate grade point average and adjusted compensation for both line and staff jobs directly refutes the claim made by Livingston that there is no association between performance in a graduate management program and managerial success. Graduate grades have a

TABLE 2

REGRESSION RESULTS ON (LINE) MANAGERIAL COMPENSATION

Variables	$\bar{X}$	*SD*	(1)	(2)	(3)	Zero-order correlation with adjusted compensation (on entire sample)
			Beta Weights (on entire sample)			
			With 1969 compensation as dependent variable	With 1969 compensation adjusted for years of work experience	Same as (2) but without using graduate GPA as predictor	
Dummy variable (8–9 years work experience)	.21	.70	.21[d]	—	—	—
Dummy variable (6–7 years work experience)	.15	.67	.27[d]	—	—	—
Dummy variable (4–5 years work experience)	.00	.57	−.11[d]	—	—	—
Graduate GPA	6.14	.72	.49***	.51***	—	.49***
Undergraduate (GPA) science	2.69	.81	−.38***	−.39***	−.33**	−.09
Mathematics and statistics major	.08	.28	.26**	.27**	.24**	.25**
Rating of involvement in sports activities	1.19	1.74	.14	.15	.28**	.21*
Undergraduate GPA (Junior)	2.88	.58	.16	.16	.40***	.16
Current Compensation ($)	21,169[a]	5,707[a]	—	—	—	—
df	—	—	53	56	57	60
R[b]	—	—	.68***	.64***	.46***	—
(Double) Cross-validated *R*[c]	—	—	.57***	.52***	.28**	—

Note. Columns designated (1), (2), and (3) summarize the results of the stepwise regressions. GPA = grade point average.

[a] After adjusting for the number of years of work experience (using the regression coefficients for the dummy variables), the new mean and standard deviation were 15240 and 5412, respectively.

[b] See under Table 1.

[c] See under Table 1.

[d] The estimation of parameters is done in such a way that the regression coefficient for the dummy variable for 3 years work experience is the negative of the sum of the regression coefficients for these three dummy variables (the years of work experience reported here are after graduation from GSIA).

* $p < .10$.
** $p < .05$.
*** $p < .01$.

greater effect on line jobs than staff jobs which may, in part, be due to the reduced variability of adjusted compensation in staff compared to line jobs. The coefficient of variation of adjusted compensation was .24 and .36 for staff jobs and line jobs, respectively.

A related finding was the pronounced effect of work experience on staff jobs compared to line jobs. The reduction in standard deviation due to adjusting compensation for work experience was $1,039.00 and $295.00 for staff jobs and line jobs, respectively. One explanation for this difference is that staff work is more difficult to evaluate with respect to organizational performance, allowing tenure to play a more important role in the reward system for staff jobs compared to line jobs.

The significance of the early identification models based on academic and extracurricular activity variables support the ability to predict managerial success prior to entry into graduate school. With the exception of the mathematics and statistics major, the variables that enter the staff model differ from those that enter the line model. The reasons for these differences may be a function of differential abilities necessary for success and/or self-selection, that is, different types of persons entering these two career patterns.

More research needs to be conducted to test these hypotheses.

Although in a previous study, Srinivasan and Weinstein (1973) found seven significant predictors of graduate grades at the Graduate School of Industrial Administration of Carnegie-Mellon University, the present study indicated that only two of the predictors of adjusted compensation (junior grades and engineering major) predict graduate grades. This finding suggests that schools that solely use academic criteria for selection may be ignoring variables which predict future career performance. Management schools are justified in searching for bright graduate students, but if they are at all interested in their ability to produce high-earning (successful) graduates, they should consider alternative selection strategies.

The results reported in this study were somewhat restricted to business school graduates from analytically-oriented programs similar to that of the Graduate School of Industrial Administration of Carnegie-Mellon University. In order to generalize these results to the whole population of MBA graduates, it is necessary to study a broader sample including students from other business schools. Furthermore, to directly answer the question of the relevance of business education to managerial success, one should also include a control group of managers who did not have a formal management education. Finally, the results of the present study are based on students who were admitted and who subsequently enrolled in the MBA program. The predicted relationship may not be valid for the applicant population as a whole. The results must be corrected for "curtailment" or "restriction of range" (Thorndike, 1949) before they can be used in a valid model for selecting students from an entire population of applicants.

Three major conclusions can be derived from this study: (*a*) Graduate grade point average is significantly correlated with management compensation 3–10 years after graduation. (*b*) The effect of work experience on annual compensation is considerably stronger for staff managers than for line managers 3–10 years after graduation. (*c*) Staff and line management compensation can be differentially predicted from pregraduate and graduate data.

REFERENCES

BRENNER, M. H., & LOCKWOOD, H. C. Salary as a predictor of salary: A 20-year study. *Journal of Applied Psychology,* 1965, **49**, 295–298.

CAMPBELL, J. P., DUNNETTE, M. D., LAWLER, E. E., & WEICK, E. E. *Managerial behavior, performance and effectiveness.* New York: McGraw-Hill, 1970.

COHEN, J. Multiple regression as a general data-analytic system. *Psychological Bulletin,* 1968, **70**, 426–443.

HAIRE, M., GHISELLI, E. E., & GORDON, M. E. A psychological study of pay. *Journal of Applied Psychology,* 1967, **51**(4, Pt. 2).

HARRELL, T. W. High Earning MBA's. *Personnel Psychology,* 1972, **25**, 523–530.

LAURENT, H. Cross-cultural cross-validation of empirically validated test. *Journal of Applied Psychology,* 1970, **54**, 417–423.

LIVINGSTON, J. S. The myth of the well-educated manager. *Harvard Business Review,* 1971, No. 1, **49**, 79–89.

MARSHALL, G. L. Predicting executive achievement. Unpublished doctoral dissertation, Boston, Mass.: Harvard Business School, 1964.

NUNNALLY, J. C. *Psychometric Theory.* New York: McGraw-Hill, 1967.

SRINIVASAN, V., SHOCKER, A. D., & WEINSTEIN, A. G. Measurement of a composite criterion of managerial success. *Organizational Behavior and Human Performance,* 1973, **9**, 147–167.

SRINIVASAN, V., & WEINSTEIN, A. G. Effects of curtailment on an admissions model for a graduate management program. *Journal of Applied Psychology,* 1973, **58**, 339–346.

THORNDIKE, R. L. *Personnel Selection.* New York: Wiley, 1949.

WILLIAMS, F. J., & HARRELL, T. W. Predicting success in business. *Journal of Applied Psychology,* 1964, **48**, 164–167.

(Received January 29, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 2, 213–216

PUPILLOGRAPHY IN INDUSTRIAL FATIGUE ASSESSMENT[1]

TAMARA GEACINTOV AND W. SCOTT PEAVLER[2]

Bell Telephone Laboratories, New Brunswick, New Jersey

A study was conducted to assess pupil constriction as a reflection of fatigue in an industrial environment. Subjects were telephone operators who were participating in a trial of an automated retrieval and display device (microfilm) for directory assistance operation. Eleven subjects were tested before and after working on the traditional telephone book as well as on the mechanical device. The testing consisted of photographing the subjects' pupil at a rate of two frames per second for three minutes. A significant difference was found in the mean pupil diameter between pre- and postwork measures across work conditions. Results indicated that the technique may be useful for determining relative fatigue levels.

Objective measurement of fatigue has long evaded precise specification (Bartley, 1957). One factor which has contributed to the problem has been the notable absence of an acceptable definition of fatigue that has a specific scientific meaning (McFarland, 1971). Fatigue has traditionally been inferred by subjective reports of tiredness and/or by measurement of the presumed consequences of fatigue such as performance decrement (Przystanowicz, 1970).

Attempts to identify objective measures of fatigue have included a variety of psychological and physiological variables that range from reaction time, blink rate, Critical Flicker Frequency (CFF), and visual acuity to biochemical, circulatory, and psychophysiological analyses of bodily processes. The major difficulty with many of these techniques is that the human organism is able to temporarily overcome effects of fatigue—except in cases of extreme exhaustion—during a test situation, thereby offsetting all or some of the dependent variables (Cameron, 1971; Dearnaley, 1958).

An area of investigation which has been even more elusive is the measurement of fatigue accruing as a result of normal work activity at jobs that do not require an inordinate amount of physical activity. This mental/psychological fatigue is presumably a normal result of repetitive and/or monotonous tasks and may be partially caused by boredom.

One physiological system which under controlled conditions seems to yield a rather sensitive indication of central and/or mental activity patterns is the pupillary apparatus (Goldwater, 1972). In a series of studies that investigated pupillary dilation patterns associated with psychological processes (Kahneman & Peavler, 1969; Kahneman, Peavler, & Onuska, 1968; Peavler, 1969; Peavler & McLaughlin, 1967), a consistent observation was a reliable decrease in baseline diameter between the onset and termination of an experimental session, regardless of the nature of the task, which suggested that systematic diameter changes may be reflective of normal cognitive effort associated with typical job activities.

Other investigators of the pupillary system (Lowenstein & Loewenfeld, 1951) have indicated a close relationship between pupillary patterns and fatigue levels. They state that pupillary waves of dilation and contraction accompany arousal and drowsiness in the tired individual and that the waves are ideal indicators of fatigue because of their objective, quantitative nature and sensitivity (Lowenstein, Feinberg, & Loewenfeld, 1963). Under normal light conditions, pupil diameter is maximum in a well-rested and alert subject and diminishes as fatigue increases—finally reaching a minimum diameter just prior to the onset of sleep (Lowenstein & Loewenfeld, 1964).

---

[1] Certain aspects of the technique in this study have been previously reported at symposia on the pupil held by the German Ophthalmological Society, Bad Nauheim, Germany, March 1972 and the American Psychological Association, Honolulu, Hawaii, September 1972.

[2] Requests for reprints should be sent to W. Scott Peavler, P.O. Box 2020, Bell Telephone Laboratories, New Brunswick, New Jersey 08903.

A review by Bell Laboratories of the directory assistance operation, in order to identify labor or timesaving methods and devices, resulted in a trial examination of new systems for storage and retrieval of customer listings. One of the proposed systems consisted of a microfilm viewer that enabled the operator to access the telephone directory listings using an electronic keyboard. Customer listings were displayed on a screen prior to a final search. Because the operator works many hours per day at this task, concern was expressed that continuous usage of display devices could possibly induce undue fatigue on the part of the operator.

As a result of the reviewed evidence relating pupil size to fatigue, a pilot study employing 10 subjects was executed. Presumed fatigue was induced by exposing the subject to three Hollywood films, all shown on one day, with a total viewing time of six hours interrupted by a lunch period and a short afternoon break. Promising results of that study (reliable decreases in baseline diameter were observed) resulted in plans to obtain relative measures from operators participating in the microfilm trial while working at both the traditional book positions as well as the microfilm positions.

Previous work by Lowenstein and his associates that related pupil size to fatigue level was confined primarily to clinical evaluations of alert subjects and chronically tired individuals, whereas the purpose of the present study was to determine if the pupillary response sensitivity permits discrimination between two similar job activities.

METHOD

Subjects and Apparatus

Seventeen operators participated in the study. Measurements on two of the subjects were not completed due to operator schedule changes. The data of four additional subjects were discarded because the pupils were unmeasurable for various reasons. As a result, the data were based on 11 subjects, 5 of whom normally used corrective lenses that were not used during data collection periods.

The pupillary data were obtained with the technique developed by Hess (1972). The apparatus consisted of a viewing cabinet with a mounted camera and red source light that permitted infrared photographic records of pupil movements to be recorded. Measurements were taken with the room in darkness except for light from a desk lamp on the rearview screen of the viewing box. The subjects looked straight ahead at a set of three letters (ABC) after having been instructed to focus on the middle letter (B) located about one third of the distance from the top of the rearview screen.

Procedure

Each subject's pupil was allowed to adjust to the red light for one minute, after which it was photographed for three minutes at a rate of two frames per second during each of the four sessions, and resulted in 360 measures at each session. Each subject was measured before and after a full day's work on the book and before and after a full day's work on the microfilm reader. The observations were taken on two consecutive days and resulted in four sets of measures for each subject. Half of the subjects were initially measured before and after working on the book position, while the other half were measured initially before and after working on the machine positions. Each frame of the processed film was projected on a screen 89 centimeters from the front of the lens of a Dunning Animatic single-frame projector, and the pupil image, magnified 15 times, was measured to the nearest millimeter.

RESULTS AND DISCUSSION

The pupil is sensitive to many influences other than fatigue. Other variables which affect pupil size include psychic stimulation, mental concentration, anxiety, attention, motivation, emotional excitement, etc. In order to isolate the effect of fatigue, an attempt was made to minimize influences of extraneous variables by photographing the subject's pupil in a dark, quiet room, with only the experimenter present. Also, because cognitive relaxation is necessary in order to obtain data uncontaminated by the factors mentioned, data analysis was limited to the final minute of the recording period. It should be added that continuation of the session in a partially sensory-deprived environment probably may have resulted in some subjects going to sleep.

The mean pupil sizes for all subjects before and after work on the book position and the microfilm position are shown in Figure 1. The graph encompasses the entire three-minute recording period. For easier graphic presentation, data points have been averaged for every 15 seconds.

The initial mean pupil size for all the subjects at the beginning of the workday was essentially the same at the book and machine positions: book—4.6 millimeters, microfilm—4.5 millimeters. Furthermore, individual pupil sizes obtained during the morning session were not correlated with diameter differences between morning and afternoon sessions.

The mean difference in pupil size (.32 millimeter) between the morning and the evening measures for all 11 subjects across work conditions was statistically significant ($t = 2.78$, $p < .01$). Further analysis of the data within each work condition indicated that the mean decrease (.43 millimeter) was statistically significant ($t = 2.64$, $p < .05$) when the subjects worked on the microfilm device but not when they worked on the traditional book positions.

The purpose of the study within which the fatigue measures were carried out was to compare the performance of operators who worked on the microfilm machine with that of operators who used the paper telephone record. This principal study showed that using the mechanized device resulted in improved performance by operators because of the inherently greater access speed of the mechanized microfilm device. Therefore, performance changes traditionally correlated with fatigue were confounded with the differences in treatment conditions.

While subjective reports of fatigue from the specific operators participating in the study were not identifiable because of guarantees of anonymity, the population from which they were selected reported significantly more physical symptoms of fatigue such as neckaches, backaches, eyestrain, and headaches when working on the microfilm device than when working at the paper positions. These complaints were somewhat contradicted, however, by an overwhelming preference for the microfilm device.

A difference in relative fatigue indexes was not unanticipated, because several plausible contributing factors associated with the microfilm device had been identified both prior to and during the study. The major factors, identified both by operator comment and by environmental analysis, included poor overhead lighting that created glare, excessive contrasts in the visual field, specular reflections, inappropriate chair design that did not provide proper support, and an inadequate physical design of the microfilm device that precluded shifts into restful positions. It should be noted that operator acceptance of the device, as evaluated, was generally very positive and enthusiastic in spite of the noted problems.

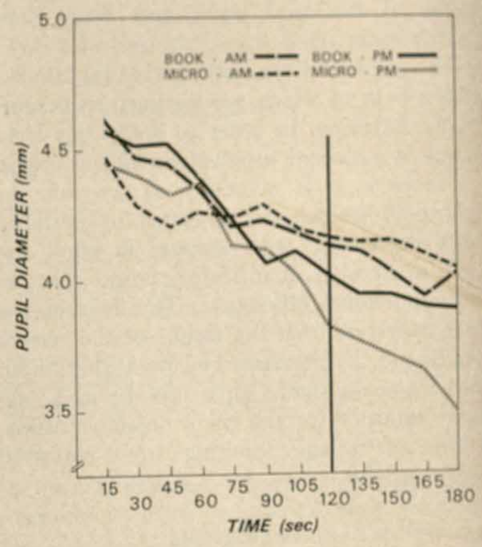


FIGURE 1. Mean pupil-diameter changes for 11 subjects across time before and after working on book and microfilm positions.

The interpretation of fatigue in the present study rests primarily with the subjective reports, the environmental factors that were identified as selectively affecting the microfilm operation, and the existing literature. Reliable performance differences, even without the previously mentioned confounding between tasks, are extremely difficult to obtain in the live directory assistance environment because of the nature of the task. Since the customer, who is not under experimental control, is the most potent determiner of speed and accuracy in the transaction, data collection must necessarily continue over weeks or even months before reliable individual differences in operator performance may be identified.

In the present study, we were able only to speak of statistical significance when comparing mean pupil size at the end of a day's work performed on two somewhat similar tasks, both of which are assumed to be normally fatiguing. In order to apply this technique to practical problems involving groups of subjects, it is necessary to correlate the degree of fatigue, as measured by performance decrement, with changes in pupil size and to develop quantitative norms that reflect individual differences. The investigators feel, however, from the results of the present study and the literature reviewed, that pupillary response shows some promise as a reliable indicator for the more objective assessments of fatigue resulting from industrial work activity.

REFERENCES

BARTLEY, S. H. Fatigue and inadequacy. *Physiological Review,* 1957, **37**, 301–324.

CAMERON, C. Fatigue problems in modern industry. *Ergonomics,* 1971, **14**, 713–720.

DEARNALEY, E. J. Fatigue: A psychological review. (I. A. M. Report 110) London: Royal Air Force Institute of Aviation Medicine, 1958.

GOLDWATER, B. C. Psychological significance of pupillary movements. *Psychological Bulletin,* 1972, **77**, 340–355.

HESS, E. H. Pupillometrics. In N. S. Greenfield & R. A. Sternbach (Eds.), *Handbook of psychophysiology*. New York: Holt, Rinehart & Winston, 1972.

KAHNEMAN, D., & PEAVLER, W. S. Incentive effects and pupillary changes in association learning. *Journal of Experimental Psychology,* 1969, **70**, 312–318.

KAHNEMAN, D., PEAVLER, W. S., & ONUSKA, L. Effects of verbalization and incentive on the pupil response to mental activity. *Canadian Journal of Psychology,* 1968, **22**, 186–196.

LOWENSTEIN, O., FEINBERG, R., & LOEWENFELD, I. E. Pupillary movements during acute and chronic fatigue. *Investigative Ophthalmology,* 1963, **2**, 138–157.

LOWENSTEIN, O., & LOEWENFELD, I. E. Types of central autonomic innervation and fatigue. *Archives of Neurology and Psychiatry,* 1951, **66**, 581–599.

LOWENSTEIN, O., & LOEWENFELD, I. E. The sleep-waking cycle and pupillary activity. *Annals of the New York Academy of Sciences,* 1964, **117**, 142–156.

MCFARLAND, R. A. Understanding fatigue in modern life. *Ergonomics,* 1971, **14**, 1–10.

PEAVLER, W. S. Attention, processing load, and pupil size. Paper presented at the Sixth Colloquium on the Pupil, Bethesda, Maryland, May 1969.

PEAVLER, W. S., & MCLAUGHLIN, J. F. The question of stimulus content and pupil size. *Psychonomic Science,* 1967, 8(12), 505–506.

PRZYSTANOWICZ, A. Effect of fatigue on the precision of positioning movements. *Przegląd Psychologiczny,* 1970, **19**, 33–44.

(Received January 8, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 2, 217–218

SHORT NOTES

GETTING THE MAIL QUESTIONNAIRE RETURNED: SOME PRACTICAL RESEARCH CONSIDERATIONS

JOHN F. VEIGA[1]

University of Connecticut

Methods used for returned mail questionnaires have not been systematically evaluated in terms of their effect on the rate of return and respondent bias. To evaluate the effectiveness of three most often used return methods, 100 managers received a questionnaire which was to be returned by either a stamped envelope, a business reply envelope or their firm's interplant mail system. Results suggest that the interplant mail system provided the greatest return rate, at lowest cost, without respondent bias. The business reply envelope produced a very poor response rate.

Even though the choice of a return vehicle for a mail questionnaire can have a substantial impact on research, rarely have return vehicles ever been tested as rigorously as the data which they help generate. As part of a major research project, a study was conducted to evaluate three return methods: (*a*) the preaddressed stamped envelope, (*b*) the postage guaranteed business reply envelope, and (*c*) an interplant (in-house) mail system. The stamped envelope, while being the most expensive in terms of postage cost, requires the respondent to waste a postage stamp suggesting a possible advantage. The business reply envelope avoids the added expense of unreturned questionnaires but might be associated with junk mail. The interplant method has the advantage of saving postage cost but it could create substantial bias if the respondents feel their anonymity might be jeopardized.

METHOD

Return Rates

One hundred randomly selected managers received a questionnaire. These managers were divided into three groups depending on the type of return mailer received. Table 1 presents the initial results.

In observing the return rate, it appears that the poor response from the business reply offers some support for the junk mail hypothesis. In addition, the findings suggest that the other two methods are superior to the business reply and are about equal in return effectiveness. Also of interest was the length of time required to get the mailers returned. It appears that the questionnaires returned by business reply somehow received a lower priority in getting filled out and returned than either of the other methods. To get a closer look at the length of time required, weekly return rates were recorded for four weeks after the study began and are reported in Table 2.

The data in Table 2 further substantiate the ineffectiveness of the business reply mailer. More important is the large discrepancy noted in the first week returns between the interplant system and the stamped envelope. Two explanations for the discrepancy can be posited. First it could be argued that the questionnaires returned by the interplant system received attention faster because the return method demonstrated evidence of company support. Second, it was noted that about 20% of the stamped envelopes were from distant geographic locations. Thus it appears that stamped envelopes provided the respondents with an opportunity to take the questionnaire home or on a business trip thereby

TABLE 1

QUESTIONNAIRE RETURNS BY MAILER TYPE

Mailer type	Questionnaire returns	
	Return rate (%)	Length of return period (in days)
Interplant	82	13
Stamped envelope	80	17
Business reply	57	25

---

[1] Requests for reprints should be sent to John F. Veiga, Department of Industrial Administration, University of Connecticut, Storrs, Connecticut 06268.

TABLE 2

WEEKLY RETURN RATE BY MAILER TYPE

Mailer type	% received within			
	1st week	2nd week	3rd week	4th week
Interplant	72	28	—	—
Stamped envelope	39	54	7	—
Business reply	44	22	22	12
Total	53	36	8	3

Note. In order to overcome the United States mail-processing time, the data reported reflect the postmarked date as the receipt date. Moreover, since 80% of all questionnaires returned by United States mail were mailed within the firms sampled, this adjustment clearly eliminates United States mail-processing time and allows for comparative analysis of respondent reaction times.

increasing return time. It should also be noted that 89% of all questionnaires returned were returned within two weeks indicating that two weeks is a sufficient length of time for go/no-go decisions on second mailings and pilot study projections.

What About Bias?

It seems fairly obvious that the interplant mail system is the superior return method at least in terms of cost and return rate. However, as indicated above, part of the efficiency of the interplant method was attributed to the implied company support. Hence, the concern over potential respondent bias needs to be dealt with before the interplant method can be considered.

In order to assess potential respondent bias, three of the 13 attitudinal measures used in the questionnaire were randomly selected for analysis. These three measures included: (*a*) salary satisfaction, (*b*) supervisory satisfaction, and (*c*) satisfaction with company policies. The questionnaires returned by the stamped envelope provided the unbiased control group for a comparative analysis. It was hypothesized that since the sample was randomly selected, the data generated on the attitudinal questions should have statistically similar distributions regardless of the mailer used. The Mann–Whitney U-test was employed to test the hypothesis. U-values were all nonsignificant at .01 level of significance and, thus, it was concluded that the interplant system did not produce any significant respondent bias.

SUMMARY

The study supports the use of the interplant system as the most efficient and economical mail return method. Contrary to what had been hypothesized, the interplant method did not produce any respondent bias. In a cost conscious research environment, it provides a return vehicle that could save a considerable amount of money in projects where the sample size is quite high. Moreover, it is clear that the business reply mailer is not an effective return method while the stamped envelope could be a second choice if an interplant or in-house system is not available. Certainly one could argue that several other factors affect mail returns such as, length of questionnaire, profile of individuals sampled, etc. Mainly, the system finally selected depends on the trade offs the researcher makes.

(Received November 8, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 2, 219-221

EFFECTS OF ALTERNATIVE FOLLOW-UP PROCEDURES ON MAIL SURVEY RESPONSE RATES[1]

MICHAEL J. ETZEL[2] AND BRUCE J. WALKER

University of Kentucky

In a large-scale mail survey of consumers, this experiment compared the effects on response rate of (*a*) not sending a follow-up letter, (*b*) sending a follow-up letter with a duplicate questionnaire and return envelope, and (*c*) sending a follow-up letter without duplicates. The results suggest that the best method is a follow-up without duplicates.

Among the problems in the use of mail questionnaires, one of the most vexing is low response rate. Response rates between 10% and 30% are common (Boyd & Westfall, 1972; Luck, Wales, & Taylor, 1970).

Previous research on response rates can be grouped into three categories, according to when the variable under consideration is manipulated. Thus, efforts to stimulate response rates may come in the following three categories: (*a*) prior to mailing the questionnaire, for example, preliminary phone calls or letters (Ford, 1967; Heaton, 1965; Kephart & Bressler, 1958; Myers & Haug, 1969; Parsons & Medford, 1972; Stafford, 1966); (*b*) coincidental to mailing the questionnaire, for example, type of postage (Kephart & Bressler, 1958; Kernan, 1971; Watson, 1965), form of address (Kernan, 1971), type of signature in cover letter (Kawash & Aleamoni, 1971), inclusion of incentive (Kephart & Bressler, 1958; Watson, 1965; Wotruba, 1966), nature of cover letter (Andreasen, 1970; Simon, 1967), and questionnaire format (Brown, 1965–1966; Ford, 1968); and (*c*) after mailing the questionnaire, for example, follow-up interviews, phone calls, and correspondence (Francel, 1966; Hochstim & Athanasopoulos, 1970; Myers & Haug, 1969; Nichols & Meyer, 1966–1967; Watson, 1965). The present study is related to the third category.

It is generally accepted that one or more follow-ups increase the rate of response (Francel, 1966). However, recent findings regarding this factor are inconclusive. For example, in a study of "bank image," a follow-up letter one week after the initial letter produced no greater overall return than no follow-up letter (Myers & Haug, 1969).

On the other hand, in a study of *Business Week* subscribers, a two-day follow-up mailing was effective in increasing response rate (Watson, 1965). The control group responded at a 30% level compared to 46% for the group that also received a follow-up mailing.

Furthermore, a recent study by Dillman (1972) examined the response rate effects of including a duplicate questionnaire in a follow-up mailing. However, the results were inconclusive because of methodological problems.

It is apparent, then, that the effects on response rates of alternative follow-up methods have not been established empirically. Therefore, the present study investigated response rates produced by three mail survey procedures: (*a*) not sending a follow-up letter, (*b*) sending a follow-up with a duplicate questionnaire and return envelope, and (*c*) sending a follow-up letter without duplicates. The following null hypothesis was tested: The response rates produced by three different follow-up procedures do not differ significantly.

METHOD

An experimental study was conducted in conjunction with a study of consumers' credit purchasing habits and buying preferences. The sample for the credit study consisted of 1,500 names randomly drawn from a list of bank credit card holders provided by a bank. For the present study, a subsample of 700 names (male and female adults) was drawn from the 1,500 names.

The questionnaire used in the study was four pages long and required about 15 minutes to complete. Only "check-type" answers were required. The responses were anonymous.

The study was conducted under typical survey research conditions. The sample was large and representative of the general population in the study area. Additionally, the topic investigated in the survey (bank credit card usage) and the questionnaire itself were not atypical in any way.

---

[1] The authors wish to acknowledge the financial support provided by the Kentucky Research Foundation for this study.

[2] Requests for reprints should be sent to Michael J. Etzel, Department of Business Administration, University of Kentucky, Lexington, Kentucky 40506.

TABLE 1

MAIL SURVEY RESPONSE RATES FROM THREE DIFFERENT SAMPLE GROUPS

Sample	Original base	Revised base[a]	No. responses	Net % response rate
Group 1—control (no follow-up)	300	299	116	38.8
Group 2—experimental (follow-up without duplicate questionnaire and return envelope)	200	200	118	59.0
Group 3—experimental (follow-up with duplicates)	200	193	102	52.9

[a] The revised base reflects adjustments due to questionnaires returned as undeliverable.

Thus, the study possesses a high degree of external validity; as pointed out (Luck, Wales, & Taylor, 1970, p. 121), the conclusion can properly be projected, therefore, to the real-world problem of the decision maker.

The subsample of 700 was divided into three groups: a control group of 300 (Group 1) and two experimental groups of 200 each (Groups 2 and 3). Each sample member was sent the questionnaire packet (an individually addressed, personally signed cover letter explaining the credit survey; a questionnaire; and a postage-paid, self-addressed return envelope).

Five days after the initial mailing, the experimental groups were sent follow-up letters.[3] Group 2 received another individually addressed, personally signed letter requesting cooperation. Group 3 received the same follow-up letter *and* a duplicate questionnaire and return envelope. The letters were identical except for a necessary addition in the Group 3 letter: "In case you have misplaced either the questionnaire or the postage-paid envelope, duplicates are enclosed."

RESULTS

The cut-off date for examining response rate was 17 days after receipt of the first completed questionnaire. Only four questionnaires were received after this date. To that date, 336 completed questionnaires had been received for an overall response rate of 48.6%.

The response rates for the groups are shown in Table 1. The control group response rate of 38.8% (116 returned from a revised base of 299) was higher than normal for a mail questionnaire, particularly a four-page one. Perhaps this response rate reflected an interest in the subject matter among the respondents and the relative ease with which the instrument could be completed. The response rates for both experimental groups, 59.0% and 52.9%, respectively, were considerably higher than the control group.

The chi-square analysis used to test the hypothesis indicated that both groups receiving follow-up correspondence (Groups 2 and 3) exhibited significantly higher response rates than the group receiving only the initial mailing. Comparing Group 2 and the control group, $\chi^2 = 19.6$, $p < .001$; comparing Group 3 and the control group, $\chi^2 = 9.4$, $p < .001$.

The difference in response rates between Groups 2 and 3, those that received the alternative follow-ups, was not significant ($\chi^2 = 1.5$, $p > .05$). Thus, including a duplicate questionnaire and return envelope in the follow-up did not produce a greater proportion of responses than did sending the follow-up without duplicates.

According to the results of this study, the researcher who uses a mail questionnaire and wishes to maximize response rate should consider employing a follow-up letter *without* duplicates of the questionnaire and return envelope. This procedure not only should improve the response rate to a greater extent than that generated by a single mailing but also should avoid the additional printing, clerical, and perhaps postage expenses connected with including duplicates in a follow-up letter.

REFERENCES

ANDREASEN, A. R. Personalizing mail questionnaire correspondence. *Public Opinion Quarterly,* 1970, **34**, 273–277.

BOYD, H. W., JR., & WESTFALL, R. *Marketing research.* (3rd ed.) Homewood, Ill.: Richard D. Irwin, 1972.

BROWN, M. L. Use of a postcard query in mail surveys. *Public Opinion Quarterly,* 1965–1966, **29**, 635–637.

DILLMAN, D. A. Increasing mail questionnaire response in large samples of the general public. *Public Opinion Quarterly,* 1972, **36**, 254–257.

FORD, N. M. The advance letter in mail surveys. *Journal of Marketing Research,* 1967, **4**, 202–204.

FORD, N. M. Questionnaire appearance and response rates in mail surveys. *Journal of Advertising Research,* 1968, **8**, 43–45.

FRANCEL, E. G. Mail-administered questionnaire returns with a preliminary letter. *Journal of Marketing Research,* 1966, **3**, 89–92.

---

[3] To stimulate response from those favorably disposed or neutral regarding the questionnaire, the follow-up, which represents a reminder, should be sent while the potential respondent still recalls the nature of the survey. Thus, a relatively short interim, five days, was used.

HEATON, E. E., JR. Increasing mail questionnaire returns with a preliminary letter. *Journal of Advertising Research*, 1965, **5**, 36–39.

HOCHSTIM, J. R., & ATHANASOPOULOS, D. A. Personal follow-up in a mail survey: Its contribution and its cost. *Public Opinion Quarterly*, 1970, **34**, 69–81.

KAWASH, M. G., & ALEAMONI, L. M. Effect of personal signature on the initial rate of return of a mailed questionnaire. *Journal of Applied Psychology*, 1971, **55**, 589–592.

KEPHART, W. M., & BRESSLER, M. Increasing the responses to mail questionnaires: A research study. *Public Opinion Quarterly*, 1958, **22**, 123–132.

KERNAN, J. B. Are 'bulk rate' occupants really unresponsive? *Public Opinion Quarterly*, 1971, **35**, 421–422.

LUCK, D. J., WALES, H. G., & TAYLOR, D. A. *Marketing research.* (3rd ed.) Englewood Cliffs, N.J.: Prentice-Hall, 1970.

MYERS, J. H., & HAUG, A. F. How a preliminary letter affects mail survey returns and costs. *Journal of Advertising Research*, 1969, **9**, 37–39.

NICHOLS, R. C., & MEYER, M. A. Timing postcard follow-ups in mail questionnaire surveys. *Public Opinion Quarterly*, 1966–1967, **30**, 637–646.

PARSONS, R. J., & MEDFORD, T. S. The effect of advance notice in mail surveys of homogeneous groups. *Public Opinion Quarterly*, 1972, **36**, 258–259.

SIMON, R. Responses to personal and form letters in mail surveys. *Journal of Advertising Research*, 1967, **7**, 28–30.

STAFFORD, J. E. Influence of preliminary contact on mail returns. *Journal of Marketing Research*, 1966, **3**, 410–411.

WATSON, J. J. Improving the response rate in mail research. *Journal of Advertising Research*, 1965, **5**, 48–50.

WOTRUBA, T. R. Monetary inducements and mail questionnaire response. *Journal of Marketing Research*, 1966, **3**, 398–400.

(Received December 26, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 2, 222-225

USING MONETARY INDUCEMENTS TO INCREASE RESPONSE RATES FROM MAILED SURVEYS:

A REPLICATION AND EXTENSION OF PREVIOUS RESEARCH

SCHUYLER W. HUCK [1] AND EDWIN M. GLEASON

University of Tennessee

In a previous investigation, Kephart and Bressler demonstrated that the response rates to mailed questionnaires could be significantly increased by including a monetary incentive of 25¢. Using 200 college students, the present study indicates that (*a*) a quarter (25¢) is still an effective incentive, despite its devaluation during the past 15 years, (*b*) the Kephart and Bressler study can be generalized beyond their unique target population, and (*c*) a follow-up mailing sending the quarter to nonresponders rather than to everyone on the initial mailout saves about 50% of the cost of the financial incentive but does not significantly decrease the response rate.

One need only glance through the professional journals of many disciplines to note that the mailed questionnaire continues to be a most popular data gathering tool in research investigations. Unfortunately, the response to mailed questionnaires is typically poor. Returns of less than 50% are common, while higher percentages are rare (Helmstadter, 1970). Since nonresponders may be different from responders, a researcher cannot be certain that generalizations from his data are valid when the response rate is low. For this reason, various attempts have been made to identify factors that increase the rate of return from mailed questionnaires. Researchers have investigated the possible effect of using airmail stamps rather than regular postage on the return envelope (Orr & Neymay, 1965), of having an actual signature on the cover letter (Kawash & Aleamoni, 1971), and of sending gifts to the recipients of the questionnaire (Pucel, Nelson, & Wheeler, 1971).

As one might suspect, research has demonstrated that a monetary inducement can be used to improve response rates. In a marketing survey, Robinson and Agism (1951) found a quarter (25¢) to be effective in substantially increasing the return rate whereas lesser coins or stamps were not; a dollar bill, however, was not more effective than the quarter. Kephart and Bressler (1958) applied these findings to the educational setting and concluded that although the quarter was effective in increasing the response rate (70% versus 52%), it was no more effective than a questionnaire with an ordinary follow-up to nonresponders (68%).

The Kephart and Bressler study (1958) raises three important questions. First, inflation has caused the value of a quarter to decrease since the late 1950s,[2] and it is questionable whether 25¢ can still be used as an effective monetary incentive. Second, the target population surveyed by Kephart and Bressler was, by their own admission, quite unique, highly motivated, professional female nurses. Assuming that their results could be replicated in spite of the lower value of a quarter, it is questionable whether the quarter could "work" in the more general academic community. Third, and most important, it seems reasonable to ask whether it is better to send the quarter on a follow-up mailing to nonresponders as compared with sending it on the initial mailout to all subjects. In the Kephart and Bressler study, many individuals receiving the monetary incentive would have returned the questionnaire even if the quarter had not been included. In fact, returns from the control group suggest that the quarter was wasted on approximately 52% of the individuals to whom it was sent. Although it obviously represents substantial monetary savings to send the quarter only to those individuals who do not initially respond, this procedure cannot be recommended unless it can be shown that the delayed use of the monetary incentive results in at least as high a response rate as does an initial use of this inducement.

The present research investigation was undertaken in an attempt to answer all three of the questions discussed above. Although answers to the first two questions were considered interesting in and of themselves, the researchers were most concerned with the third question. More specifically, the present study was designed to find out whether it is better to send a quarter (*a*) in an initial mailout to all individuals, (*b*) in a follow-up to those individuals who do not respond to the initial mailout, or (*c*) in a second follow-up mailing to those individuals who do not respond to either the initial mailout or the first follow-up mailing.

---

[1] Requests for reprints should be sent to Schuyler W. Huck, Department of Educational Psychology and Guidance, University of Tennessee, Knoxville, Tennessee 37916.

[2] In terms of purchasing power, the Consumer Price Index shows that 25¢ in 1958 was worth only 17.1¢ in October, 1972, the time of the present study.

METHOD

Subjects

During the fall academic term, 1972, 200 subjects were randomly selected from approximately [illegible] male and female students living in residence halls at the University of Tennessee. Once identified, these subjects were randomly assigned to one of four treatment groups.

The Questionnaire and Accompanying Materials

The questionnaire used in this study was a variation of the Rokeach Dogmatism Scale (Rokeach, 1960). This instrument contained the 40 statements that appeared in Form E of the original Rokeach scale, plus an additional 10 statements that had been included for the purpose of a previous research study. The directions requested the subjects to indicate the degree to which they agreed or disagreed with each statement by means of a 6-point Likert-type rating.

In addition to the questionnaire, each of the subjects received a cover letter and a return envelope. The cover letter was mimeographed and began "Dear U–T Student." The letter explained that the recipient had been randomly selected for a graduate research project and that the questionnaire should be completed and returned within a two-week interval. The cover letter was signed with the name of a fictitious graduate student. The preaddressed envelope did not require any postage as it was clearly marked for return through the university campus mail.

Procedure

The actual study involved four treatment groups ($n = 50$) and three mailouts. On the first mailout, all 200 subjects were sent the cover letter described previously, the questionnaire, and the return envelope. The material sent to all four groups was identical except that the cover letter for the first group of subjects (Group 1) had a quarter (25¢) taped to the bottom of the page. A short note next to the quarter thanked the recipient for taking time to fill out the questionnaire and invited him to use the quarter to buy a cup of coffee and a doughnut.

TABLE 1

SUMMARY OF THE RESEARCH DESIGN

Group	First mailout	Second mailout[a]	Third mailout[b]
1	Q + 25¢	Q	Q
2	Q	Q + 25¢	Q
3	Q	Q	Q + 25¢
4	Q	Q	Q

Note. Q = questionnaire.
[a] Only sent to subjects not responding to first mailout.
[b] Only sent to subjects not responding to second mailout.

The second mailout involved only those subjects who did not respond to the initial mailout. These nonresponders were sent a second copy of the Rokeach questionnaire and another envelope. A revised version of the cover letter indicated that no response had been received from that subject, and the recipient was asked to complete the enclosed questionnaire and return it in the accompanying envelope. For all subjects in the second treatment group (Group 2) that received this follow-up, a quarter was attached to the bottom of their cover sheet with the same note that had been put next to the quarter for subjects in the first treatment group on the first mailout.

The third mailout involved only those subjects who did not respond to either the first or the second mailout. These nonresponders (holdouts) were sent a third copy of the Rokeach questionnaire and another envelope. This time the cover letter indicated that the success of the study was contingent on a high percentage of returns and that the investigator would greatly appreciate the recipient taking a few minutes out of his schedule to complete and return the questionnaire. For all subjects in the third treatment group (Group 3) that received this final follow-up, a quarter was attached to the bottom of the cover sheet with the same explanatory note as had been used on the previous mailouts for treatment Groups 1 and 2.

The design of this experimental research investigation is summarized in Table 1. Of the four treatment groups, the first three may be thought of as experimental groups. Each of these groups received a quarter, and they differed only in terms of whether the quarter was included in the first, the second, or the third mailout. The fourth treatment group (Group 4) may be thought of as a control group, inasmuch as no monetary incentive was provided for this group at any point throughout the study.

RESULTS

On the basis of the first mailout, the Rokeach questionnaire was returned by 78% of the subjects in the first group (i.e., the group receiving the quarter) as compared with 50% of the subjects in the remaining three groups. A chi-square test indicated that the difference between these two percentages was significant ($\chi^2 = 10.8$, $df = 1$, $p < .001$). As expected, a comparison of the response rates for the three groups that did not receive the quarter yielded a nonsignificant value for chi-square ($p > .10$). These results demonstrate that (*a*) a quarter can still be used to increase the response rate, even though it is worth less than when it was initially shown to be effective as a monetary incentive in 1958, and (*b*) the conclusions drawn by Kephart and Bressler (1958) can be generalized beyond their unique population of female nurses to a much larger segment of the general academic community.

The combined returns from the first two mailouts indicated that Group 1 (a quarter on the first mailout) had a response rate of 92%; that Group 2 (a quarter on the second mailout) had a response rate of 85%; and that Groups 3 and 4 (no quarter) had an average response rate of 65%. Separate chi-square tests indicated that there was no significant difference between the percentage of returns from Groups 1 and 2, while the combined response rate from Groups 3 and 4 was significantly lower than the response rate from either of the two groups receiving the quarter. In spite of the fact that twice as many quarters had been invested in Group 1 as compared with Group 2, the response rates from these groups, after two mailouts, were not significantly different.

The combined results from all three mailouts indicated that Groups 1 through 4 had total response rates of 94%, 92%, 78%, and 71%, respectively. Chi-square tests revealed that Groups 1 and 2 did not differ in terms of these overall response rates, nor did Groups 3 and 4 differ from each other. Thus, the quarter, when delayed until the third mailout so as to minimize the researcher's financial outlay, did not seem to work. To verify this apparent finding, Groups 3 and 4 were compared in terms of the percentage of returns from only the third mailout. No significant difference was found.

DISCUSSION

On the basis of their results, Kephart and Bressler (1958) recommended sending a follow-up mailout to nonresponders rather than sending a quarter on the first mailout. This recommendation was made because the latter procedure, although not significantly different in response rate, was said to be more costly to the researcher. The present authors maintain that neither of these procedures should be adopted.

According to Kerlinger (1964), a response rate of 80%–90% or higher must be achieved before a researcher can feel confident that his results are valid. Sending a quarter on the first mailout created a response rate of 70% in the Kephart and Bressler study and 78% in the present study. On the other hand, the total response rate from an original mailout plus a follow-up mailing to nonresponders, with no monetary incentive at all, was found to be 68% in the Kephart and Bressler study and 65% in the present study. Thus, the results from both studies indicate that the "Kerlinger criterion" cannot be achieved (*a*) in one mailout by sending a quarter to everyone, or (*b*) in two mailouts by sending a follow-up mailing to nonresponders, with no monetary incentive given to anyone.

The results of the present investigation demonstrate that the monetary incentive must be combined with a follow-up in order to achieve an acceptable response rate. The researcher seems to have two alternatives: (*a*) He can send the quarter to all subjects on an initial mailout, with one or two follow-ups to nonresponders; or (*b*) he can send the quarter only to those individuals who do not respond to the initial mailout, possibly with a second follow-up sent to individuals who do not respond to either of the first two mailouts. These procedures do not significantly differ from one another in their effectiveness, and both produce a response rate which closely approaches or exceeds the criterion of 80%–90%. However, since the first procedure costs the researcher about twice as much as the second, the present authors recommend using the monetary incentive in the first follow-up, not in the initial mailout.

REFERENCES

HELMSTADTER, G. C. *Research concepts in human behavior.* New York: Appleton-Century-Crofts, 1970.

KAWASH, M., & ALEAMONI, L. Effect of personal signature on the initial rate of return of a mailed questionnaire. *Journal of Applied Psychology,* 1971, **55**, 589–592.

Kephart, W., & Bressler, M. Increasing the response to mail questionnaires: A research study. *Public Opinion Quarterly*, 1958, **22**, 123–132.

Kerlinger, F. N. *Foundations of behavioral research.* New York: Holt, Rinehart & Winston, 1964.

Orr, D. B., & Neyman, C. Considerations, costs and returns in a large-scale follow-up study. *Journal of Educational Research*, 1965, **58**, 373–378.

Pucel, D., Nelson, H., & Wheeler, D. Questionnaire follow-up returns as a function of incentives and responder characteristics. *Vocational Guidance Quarterly*, 1971, **19**, 188–193.

Robinson, R., & Agism, P. Making mail surveys more reliable. *Journal of Marketing*, 1951, **15**, 415–424.

Rokeach, M. *The open and closed mind.* New York: Basic Books, 1960.

(Received December 26, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 2, 226-227

A COMPARISON OF INDIVIDUAL, SUBGROUP, AND TOTAL GROUP METHODS OF PROBLEM SOLVING[1]

THOMAS J. BOUCHARD, JR.,[2] GAIL DRAUDEN, AND JEAN BARSALOUX

University of Minnesota

Data from three previous experiments were reanalyzed in an attempt to answer the question: Given a large number of individuals to work on a problem, should they work individually and combine their output, or should they work together in a number of small groups and pool the output of the groups? In nine contrasts, the pooled number of ideas generated by individuals was larger than the pooled number of ideas generated by groups ($p < .01$; two-tailed test). For the group sizes (4–9) and procedures (various forms of brainstorming) considered, it was concluded that the pooled-individual procedure was superior to the pooled-group procedure.

Given a reasonably large number of individuals to work on a divergent thinking type problem, a manager may have his people: (*a*) work alone and pool their ideas later, (*b*) work in one large group, or (*c*) work in small subgroups and pool their ideas later. Which procedure should be used? Research suggests that procedure (*a*) is preferable to procedure (*b*). For groups of up to nine people, the performance of nominal groups (nonoverlapping ideas of *n* individuals, where *n* = number of individuals in the real group) is superior to real groups (Bouchard, 1969, 1972; Bouchard & Hare, 1970; Campbell, 1968; Dunnette, Campbell, & Jaastad, 1963; Rotter & Portugal, 1969; Taylor, Berry, & Block, 1958). It is possible, however, that procedure (*c*) is superior to procedure (*a*). The rationale for this hypothesis is that, although a subgroup may channel the participants' thinking along similar lines and restrict the number of classes of ideas explored, it may explore ideas in greater depth. In addition, each subgroup may examine different classes of ideas, so that when the nonoverlapping ideas of several groups are summed, the total may be greater than the sum of nonoverlapping ideas for an equivalent number of individuals.

This report reexamines previously collected experimental data to determine whether it is more effective to form several subgroups and pool their results, or to use the summed efforts of individuals who worked alone.

METHOD

The raw data from three previous studies, which compared nominal groups and real groups, were reanalyzed in the following way. The nonoverlapping ideas of all same-size nominal groups in a study were pooled and the nonoverlapping ideas of the corresponding real groups were pooled. Thus, Condition A in Table 1 represents the pooling of the nonoverlapping ideas of three four-man real groups and three four-man nominal groups. The total number of ideas generated in each condition was 489 and 552, respectively. The pooling of the nominal groups' nonoverlapping ideas is equivalent to pooling the nonoverlapping ideas of 12 individuals and corresponds to what we have called procedure (*a*). The pooling of the real groups' nonoverlapping ideas corresponds to procedure (*c*). Nine independent contrasts were available that varied in terms of group size, experimental conditions, and problems worked on. The data are presented in Table 1.

Of the nine contrasts, Contrasts A, B, C, and D are from Bouchard (1972, Experiment II); Contrasts E and F are taken from an unpublished replication of part of the above study, and contrasts G, H, and I are taken from Bouchard and Hare (1970). In all nine cases, the pooled individual condition (*a*) is superior to the pooled group condition (*c*). This difference is highly significant (exact probability = .004; two-tailed test).

The results clearly suggest that for the group sizes and experimental conditions sampled, it is more effective in terms of total number of ideas

---

[1] This investigation was supported by U.S. Public Health Service Grant #1 R01 HD 05600-01 to T. J. Bouchard, Jr. The report was written while the first author was at the Oregon Research Institute, supported by General Research Support Grant RR-05612 from the National Institutes of Health and Grant MH-12972 from the National Institute of Mental Health, United States Public Health Service.

[2] Requests for reprints should be sent to Thomas J. Bouchard, Jr., Elliott Hall, University of Minnesota, Minneapolis, Minnesota 55455.

TABLE 1

NINE CONTRASTS BETWEEN POOLED-GROUP AND POOLED-INDIVIDUAL CONDITIONS RECONSTRUCTED FROM THREE PREVIOUS EXPERIMENTS

Condition	Pooled group			Pooled individual
	Size of group	No. groups	Total no. ideas	Total no. ideas
A	4	3	499	552
B	4	3	385	609
C	4	3	500	585
D	4	3	480	618
E	4	4	380	582
F	4	5	472	730
G	5	4	238	305
H	7	4	200	429
I	9	4	228	461

produced to have individuals work alone and pool their ideas later, than to have them work in subgroups and pool the ideas of the subgroups. This, of course, does not preclude the use of groups for other reasons (Maier, 1967).

REFERENCES

BOUCHARD, T. J., JR. Personality, problem-solving procedure and performance in small groups. *Journal of Applied Psychology*, 1969, **53**(1, Pt. 2), 1-29.

BOUCHARD, T. J., JR. Training, motivation and personality as determinants of the effectiveness of brainstorming groups and individuals. *Journal of Applied Psychology*, 1972, **56**, 324-331.

BOUCHARD, T. J., JR., & HARE, M. Size, performance, and potential in brainstorming groups. *Journal of Applied Psychology*, 1970, **54**, 51-55.

CAMPBELL, J. Individual versus group problem-solving in an industrial sample. *Journal of Applied Psychology*, 1968, **52**, 205-210.

DUNNETTE, M. D., CAMPBELL, J., & JAASTAD, K. The effects of group participation on brainstorming effectiveness for two industrial samples. *Journal of Applied Psychology*, 1963, **47**, 30-37.

MAIER, N. R. F. Assets and liabilities in group problem-solving: The need for an integrative function. *Psychological Review*, 1967, **74**, 239-249.

ROTTER, G. S., & PORTUGAL, S. M. Group and individual effects in problem-solving. *Journal of Applied Psychology*, 1969, **53**, 338-341.

TAYLOR, D. W., BERRY, P. C., & BLOCK, C. H. Does group participation when using brainstorming facilitate or inhibit creative thinking? *Administrative Science Quarterly*, 1958, **3**, 23-47.

(Received December 7, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 2, 228–229

EXPECTANCY AS A MODERATOR OF THE RELATIONSHIP BETWEEN JOB ATTITUDES AND TURNOVER[1]

FRED DANSEREAU, Jr.[2]
State University of New York at Buffalo
JAMES CASHMAN AND GEORGE GRAEN
University of Illinois, Urbana

The hypothesis that the strength of the predictive relationship of job attitudes on turnover is moderated by the expectancy of finding a comparable job was tested separately with a sample of office workers (222) and a sample of managers (354). Results, while supporting this hypothesis, showed little enhancement of the typical relationship.

In a recent text, *Industrial and Organizational Psychology*, Korman (1971) pointed out the lack of published research on the moderating effect of the availability of comparably alternative jobs on the empirical correlation between job attitudes and terminations. Although few investigators working in the area of attitude and behavior doubt the validity of this moderating effect, documentation has not been supplied.

Perhaps the main reason for the lack of documentation of this moderating effect is that most models make the same prediction, namely, that a participant tends to leave a given organizational field when the utility (or the like) of staying in that field becomes sufficiently lower than some alternative organizational field. Even those researchers who only report the simple correlation between job attitudes and job termination implicitly assume that the attainability of comparable jobs is essentially a constant or at least a severely restricted distribution for the sample under investigation. The point on which researchers disagree is whether the subjective probability of attaining a comparable job serves to moderate this relationship. The purpose of this analysis is to test for this moderating effect using a measure of subjective probability within samples of office workers and managers within a single organization.

METHOD

The samples for this analysis were 222 unionized office workers and 354 managers from a large international corporation in one of the basic industries. These samples were assessed as part of a larger study (Graen, Dansereau, & Minami, 1972). Thirty months after the original assessment, the employment record of each participant was obtained from organizational records and the last date of employment was noted. During this 30-month period, 23% of the office workers sample and 17.2% of the manager sample had left the organization.

For purposes of this analysis, work attitudes were indexed by the Role Orientation Index (Graen et al., 1972), and the expectancy of attaining a comparable job was assessed by asking each participant the following question: "If you really looked for another job, what do you think are your chances of finding one as good or better than your present job in the near future?" The response alternatives were: (*a*) "No chance at all," (*b*) "Probably would not find one," (*c*) "Do not know," (*d*) "Probably would find one," and (*e*) "Certain to find one." This expectancy variable was trichotomized by dividing it into low (*c* and below), medium ("Probably would find one"), and high ("Certain to find one") expectancy. These cutting points used on each sample resulted in the group frequencies shown in Table 1.

One analysis appropriate to test for the moderating effect of perceived expectancy on the correlation between job attitudes and terminations is the correlation between the predictor (attitude) and the criterion (stay or leave) within each of the three expectancy groups and over all three expectancy groups. Since the nonlinear relationship between expectancy and predictors was close to zero for both samples, little truncation of attitude scores was evident within expectancy groups.

---

[1] The authors would like to thank the president and vice president of the participating organizations for their willing cooperation and to thank the subjects in this study for their contributions. The authors would like in addition to acknowledge the research assistance of Ken Alvares, Chip Johnson, Jeanne Herman, Takao Minami, Dean Orris and Tom Johnson and the critical reading of an earlier version by Bill Haga.

[2] Requests for reprints should be sent to Fred Dansereau, Jr., Department of Organization and Human Resources, School of Management, State University of New York, Buffalo, New York 14214.

TABLE 1

POINT BISERIAL CORRELATIONS BETWEEN JOB ATTITUDES AND TERMINATION WITHIN EXPECTANCY GROUPS

Sample and predictor	Expectancy of finding a comparable job						Total sample	
	Low		Medium		High			
	r	n	r	n	r	n	r	n
Office Worker								
Work itself	−.04	76	−.21	76	−.27*	70	−.18**	222
Turnover rate	18.4%		21.0%		29.9%		23.0%	
Manager								
Supervision IR	−.04	129	−.14	127	−.21*	98	−.11*	354
Turnover rate	16.3%		15.7%		20.3%		17.2%	

Note. No[illegible] coefficients of determination (eta) of expectancy on predictors were .01 with work itself and .01 with supervision IR. IR = interpersonal relations.
* $p \leq .05$, two-tailed.
** $p \leq .01$, two-tailed.

RESULTS AND DISCUSSION

The results of this analysis showed that one of the five Role Orientation Index scales reliably predicted terminations for each sample. As shown in Table 1, work (intrinsic value) predicted termination for the office workers and supervision (interpersonal relations) predicted terminations for managers. The correlations for the total samples are typical of those reported in previous investigations (Hulin, 1966; Libo, 1953; Sagi, Olmstead, & Altelsek, 1955; Waters & Roach, 1971; Webb & Hollander, 1956; Weitz & Nuckols, 1953) approximating −.20.

As shown in Table 1, the perceived expectancy of finding a comparable job moderated the correlation between attitude and termination in both samples.[3] In both samples, the pattern of observed correlations was as expected based on the constraint of an absence of a comparable alternative. Moreover, the pattern of turnover rates of the expectancy groups in both samples is compatible with the notion that people tend not to leave their present jobs if they doubt the attainability of comparable alternatives.

Although a number of alternative hypotheses are compatible with the moderating effect of perceived expectancy on the relationship of attitudes to terminations, the empirical results supply some of the documentation required to remedy the oversight mentioned by Korman (1971). Even though these results are as predicted by our various models (theoretical significance), the strength of these relationships lack the magnitude to be of much utility to the practitioner (practical validity). A more promising approach that might be considered is described and illustrated by Johnson and Graen (1973).

REFERENCES

GRAEN, G., DANSEREAU, F., JR., & MINAMI, T. Dysfunctional leadership styles. *Organizational Behavior and Human Performance*, 1972, **7**, 216–236.

HULIN, C. L. Job satisfaction in a female clerical population. *Journal of Applied Psychology*, 1966, **50**, 280–285.

JOHNSON, T., & GRAEN, G. Organizational assimilation and role rejection. *Organizational Behavior and Human Performance*, 1973, **10**, 72–87.

KORMAN, A. *Industrial and organizational psychology*. Englewood Cliffs, N.J.: Prentice Hall, 1971.

LIBO, L. M. *Measuring group cohesiveness*. Ann Arbor: University of Michigan, Research Center for Group Dynamics, Institute for Social Research, 1953.

SAGI, P. C., OLMSTEAD, D. W., & ATELSEK, F. Predicting maintenance of membership in small groups. *Journal of Abnormal Social Psychology*, 1955, **51**, 308–311.

WATERS, L. K., & ROACH, D. Relationship between job attitudes and two forms of withdrawal from the work situation. *Journal of Applied Psychology*, 1971, **55**, 92–94.

WEBB, W. E., & HOLLANDER, E. P. Comparison of three morale measures: A survey, pooled group judgments and self-evaluations. *Journal of Applied Psychology*, 1956, **40**, 17–20.

WEITZ, J., & NUCKOLS, R. C. The validity of direct and indirect questions in measuring job satisfaction. *Personnel Psychology*, 1953, **6**, 487–494.

(Received November 13, 1972)

---

[3] In order to adjust for different turnover rates within the mobility groups biserial correlations can be used. The biserial correlations for low, medium, and high expectancy groups were −.06, −.31, and −.30 for the office workers and .06, −.21, and −.30 for the managers. Adjusting the correlations for the lack of a 50–50 split within expectancy groups had little impact on the moderator relationships.

Journal of Applied Psychology
1974, Vol. 59, No. 2, 230-232

JOB EXPECTATIONS AND VALENCES OF JOB REWARDS FOR CULTURALLY DISADVANTAGED AND ADVANTAGED CLERICAL EMPLOYEES

RICHARD D. ARVEY[1]
University of Tennessee

STEPHEN J. MUSSIO
Minneapolis Civil Service

This study examined the job expectancies of a culturally advantaged and disadvantaged employee group (female civil service clerical workers) concerning whether effective job performance would lead to certain job rewards and the relative importance or valences of these rewards compared to an advantaged group. Data revealed that the disadvantaged group saw for themselves, as opposed to the advantaged group, less likelihood for effectiveness in their jobs to lead to advancement or to the use of one"s own judgment. Also, the disadvantaged group indicated that a high salary was more important to them than to the advantaged group.

It is sometimes claimed that culturally disadvantaged employees are just not motivated to perform effectively in their jobs. One explanation behind the behavioral manifestations of these disadvantaged individuals is that they perceive that working effectively in their jobs has little effect on attaining valued work outcomes or rewards. That is, culturally disadvantaged individuals may believe that doing well in their jobs does not lead to such rewards as advancement, recognition, and so forth, regardless of whether these perceptions are true or not. This explanation stems primarily from the instrumentality-valence theory formulated by Vroom (1964). An expectancy is defined as a momentary belief concerning the likelihood that a particular act is followed by a particular outcome, while valence essentially refers to the desirability or preference an individual has for a particular outcome.

The purpose of this research was to examine the expectancies of a culturally disadvantaged employee group concerning whether doing well in their job would lead to certain reward outcomes compared to a relatively advantaged group. It is hypothesized that culturally disadvantaged individuals differ, in these job expectations, from a more advantaged group.

We also have reason to suspect that culturally disadvantaged individuals value different reward outcomes than a more advantaged group. That is, the same reward outcomes have different values or valences for the two groups. Slocum and Strawsen (1971) reported a study indicating that black certified public accountants (CPA) differed from white CPAs with regard to what they felt was important in their jobs. Blacks reported significantly higher importance ratings than other CPAs for three items: (*a*) opportunity to help people, (*b*) opportunity for independent thought and action, and (*c*) feeling of self-fulfillment. Thus, another objective of this research was to examine differences between a culturally disadvantaged and an advantaged group with respect to their perceived valences or what they felt were important aspects in a job.

METHOD

Sample

The population consisted of 266 female clerical workers employed for the civil service in a large midwestern city. All but three of these individuals had completed high school. For this population a culturally disadvantaged and culturally advantaged group were identified on the basis of two kinds of information: (*a*) whether or not their fathers had completed high school and (*b*) their scores on a measure of cultural exposure—the Environmental Participation Index (Mathis, 1967). Age was partialled out of this latter variable.

The Environmental Participation Index (EPI) is a 68-item check list where the first 20 items reflect household possessions (TV, typewriter, etc.) and are meant to indicate the richness of the material home environment if such objects were in their household prior to their sixteenth birthday. The last 48 items can be thought of as activities (e.g., visited a museum, rode in a taxi, etc.). Subjects were asked to check "yes" if they had participated in such activities by the time they were 18 years old.

---

[1] Appreciation is extended to H. Dudley Dewhirst and Michael E. Gordon for their comments on an earlier version of this article.

Also at Applied Psychological Systems, Inc., Knoxville, Tennessee.

Requests for reprints should be sent to Richard D. Arvey, Department of Industrial Management, University of Tennessee, Knoxville, Tennessee 37916.

Individuals were designated as culturally disadvantaged if their fathers had not completed high school and they were below the median on the EPI total score. The number of individuals included in the advantaged subgroup was 76—65 of whom had complete data—while the number included in the disadvantaged sample was 74—64 of whom had complete data; both subgroups were drawn from the total sample of 266.

A factor analysis revealed that these two variables were, in actuality, measures of cultural advantage (Arvey & Mussio, 1973). Similarly, according to the norms presented in the EPI manual (Mathis, 1967), the raw scores obtained by the disadvantaged sample correspond to the scores interpreted to signify a disadvantaged population.

Measures

During a time when these employees were being tested for a test validation project, they were asked to complete a questionnaire containing valence and expectancy measures.

In order to obtain information concerning the desirability or valence of certain reward outcomes held by these individuals, 10 job outcomes were presented with instructions to distribute 100 points among the outcomes according to their perceived importance. The 10 reward outcomes were: making use of abilities, accomplishment, advancement, supervise others, fair company policies and practices, high salary, getting along with co-workers, praise, use own judgment, and steady and secure employment.

Subjects' perceptions of their expectancies for the outcomes were determined by asking them to indicate the probable results of effective performance according to whether they were either quite certain the job reward would not occur; probably would not occur; was not sure one way or another; probably would occur; or was quite certain the job reward would occur. Values of 1 to 5 were assigned to these response alternatives. This expectancy was determined for each of the 10 reward outcomes, resulting then in 10 separate expectancies for each individual.

The disadvantaged and advantaged groups were compared on the separate valence and expectancy measures by obtaining means and computing *t* tests.

RESULTS

Significant differences beyond the .05 level occurred between the two groups with regard to the valences for two job outcomes. The culturally disadvantaged indicated that a high salary was more important to them than to the advantaged group, whereas when considering making use of abilities, the advantaged group indicated more valence than did the disadvantaged group. Several differences emerged that were significant at the .10 level. The disadvantaged group indicated that the following outcomes were more valent to them than to the advantaged group: steady and secure employment, praise, and getting along with co-workers. In contrast, the advantaged group indicated that accomplishment was more important to them than to the disadvantaged group. The three most important outcomes for the disadvantaged group were getting along with co-workers, safe and secure employment, and a high salary, whereas the top three outcomes for the advantaged group were use of abilities, accomplishment, and safe and secure employment.

Statistically significant differences ($p < .05$) also occurred on two expectancy variables: use of own judgment and advancement. The disadvantaged employee group saw less likelihood for effective performance in their job to lead to these two outcomes compared to the more advantaged group. In fact, 50% of the disadvantaged individuals indicated that effective performance would certainly not or probably not lead to advancement, whereas 38% of the advantaged group responded in such a manner. The data also revealed that the culturally disadvantaged group had lower expectancies on 8 of the 10 measures.

DISCUSSION

The results of this study have indicated that a culturally disadvantaged group differs from a relatively more advantaged group with regard to their valences for certain job outcomes and their expectancies concerning whether effective job performance leads to certain job rewards.

The finding that the disadvantaged group had greater valence for a high salary and secure employment than had the more advantaged group was not surprising. To some extent, these data replicate the findings of Slocum and Strawsen (1971) that blacks assign more importance to lower order needs. That the culturally disadvantaged also feel that getting along with their co-workers was the most important feature in their job was somewhat surprising. Apparently social needs were particularly acute for this sample. However, one of the items in the Slocum and Strawsen (1971) study that black certified public accountants indicated was relatively more important to them than to the whites was the opportunity to help people; this also reflected a social need. The implications of the present data are that security and social needs seem to be prepotent for the disadvantaged individuals. Companies and organizations may well strive to satisfy these needs prior to enriching and enlarging jobs, giving opportunities to use abilities, etc., which are usually associated with the work itself.

Concerning the job expectation data, there was evidence that the culturally disadvantaged saw less likelihood of advancing in the organization or making use of their own judgment as a result of working effectively in their jobs than did a more advantaged group. It seems plausible that prejudicial and discriminatory organizational practices might have had a part in lowering these expectancies. Traditionally, culturally disadvantaged individuals have been blocked from advancing to supervisory and managerial jobs. Similarly, it may threaten the jobs of supervisors and managers to permit members of the culturally disadvantaged to use their own judgment in their jobs.

The present research points to the need for renewed efforts in learning more about the culturally disadvantaged and their perceptions concerning their jobs. The findings of this study should be replicated and extended to other populations that might be considered disadvantaged.

REFERENCES

ARVEY, R. D., & MUSSIO, S. J. Determining the existence of unfair test discrimination for female clerical workers. *Personnel Psychology*, 1973, **26**, 559–568.

MATHIS, H. *Environmental participation index manual*. Psychometric Studies, Southfield, Michigan, 1967.

SLOCUM, J. W., & STRAWSEN, R. H. Racial differences in job attitudes. *Experimental Publication System*, June, 1971, **12**, Ms. No. 471[illegible].

VROOM, V. *Work and motivation*. New York: Wiley, 1964.

(Received December 29, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 2, 233-235

DEVELOPMENT OF A SCALE TO ASSESS STUDENT NURSES' VIEWS OF THE PROFESSIONAL NURSING ROLE [1]

L. M. CROCKER [2]
University of Florida

B. J. BRODIE
University of Virginia

A sample of 488 nursing students and 94 instructors rated the importance of each of 112 behaviors common to practicing nurses. A final scale was developed consisting of 60 items on which students shifted toward faculty views as class rank increased. Reliability computed for a cross-validation group was $r = .89$. Significant gains in student scores occurred as class rank increased, indicating that students adopt faculty professional views in direct relation to length of training.

A major goal of educators in the health professions is to endow students with realistic views of their professional roles. Failure to develop a realistic role image is believed to be a primary cause of nonacademic dropout, poor clinical performance, and desertion from the field after graduation. For example, Katzell (1968) reported that withdrawal after the first year of nurses' training was inversely related to fulfillment of student expectations and Fox, Diamond, and Walsh (1960) and Kibrick (1958) observed that nursing students who were realistic in their appraisals of the field were more satisfied and more likely to persist in the profession. The process by which students acquire appropriate perceptions of the professional role has been called "professional socialization."

Paramount in professional socialization theory is the axiom that faculty strongly influence student perceptions of their professions. Several researchers have attempted to measure this influence on medical and nursing students, with varying degrees of success. Pollack and Michael (1966) reported that medical students' views of the doctor image and doctor–patient relationship became more like those of their faculty as students progressed in their training. Unfortunately the statistical significance of these student response shifts were not tested. Similarly, Davis and Oleson (1964) and Oleson and Davis (1966) found that student nurses' value rating of traits were influenced by the emphasis placed on these traits by their instructors; but Siegel (1968), using the same trait list, found no significant increase in the correspondence between student and faculty professional views as students advanced in class rank. Sharp and Anderson (1972) found increasing correspondence between student and faculty scores on several subscales of the Adjective Check List as students advanced in class rank. Thus far, however, no adequate standardized instrument has been developed for measuring changes in students' views of the nursing role.

With the growing interest in reducing attrition in health care education and in learning how faculty influence student attitudes, there is a need for instruments which can be used to study the process of professional socialization. The primary objective of this study was to construct a homogeneous scale which could measure the congruence between student nurses' perceptions and faculty views of the professional nursing role.

METHOD

Instrument

An initial pool of 112 items was developed and entitled the nurses' professional orientation scale. Each item described a behavior frequently displayed by nurses in their work roles. Respondents were requested to rate each behavior on its importance for the practicing professional nurse, using a 5-point scale, ranging from extremely important (5) to undesirable (1). Sample items included such statements as: (*a*) "Questions instructions when the reason for them is not clear," and (*b*) "Accepts the death of a patient with no overt emotional signs," etc.

Scoring weights for the response to each item were determined by administering the scale to a sample of 94 nursing faculty members from the three participating universities. The proportion of faculty that endorsed a particular response was rounded to the nearest 10% and this weight

---

[1] This study was supported in part by a grant (DID NU 00232) from the Division of Nursing, U.S. Public Health Service.

[2] The authors are indebted to Isabel Payne who encouraged them to undertake this project.

Requests for reprints should be sent to Linda M. Crocker, Regional Rehabilitation Research Institute, University of Florida, 901 Lakeshore Towers, Gainesville, Florida 32601.

was assigned to that response. Thus, if a student selected a response endorsed by 90% of the faculty judges, she received 9 points for that item; if she selected a response endorsed by 20% of the faculty judges, she received only 2 points for that item, etc. Consequently, a student could achieve a high score only by rating the traits in the same way that a high proportion of faculty members had rated those traits.

Sample

The 488 student nurses who responded to the nurses' professional orientation scale were all students who could be contacted in four-year undergraduate, baccalaureate nursing programs at Michigan State University, Wayne State University, and the University of Illinois.

Design

An item analysis, cross-validation design was employed. The sample of 488 student nurses was randomly divided into two groups, stratified by class rank. The responses of the first subsample of 244 students were submitted to an item analysis. Selection of items was based on the assumption that as students progressed through the program, their views of the nursing profession become more like that of professionals. The correlation between class rank and the student's item score was tested for each item. If a significant, positive correlation was found ($\alpha \leq .05$) that item was retained for the final scale. Sixty items from the original pool of 112 were retained after the item analysis. On each of these items there was a significant shift toward agreement with faculty perception of the trait's importance as the students rose in class rank. For example, the item: "Questions instructions when the reason for them is not clear" was originally rated as extremely important by 80% of the faculty judges. When the students rated this item, that response was chosen by 28.6% of the freshmen, 50.0% of the sophomores, 83% of the juniors, and 81.8% of the seniors in the item analysis group.

The responses of the remaining 244 students in the cross-validation group were then scored for the reduced set of items selected through item analysis. The data from this independent sample were used to test the reliability and validity of this scale.

RESULTS

To test the scale's validity, the scores of the 244 students in the cross-validation group were then computed, based on the final set of 60 items. Means and standard deviations for the nurses' professional orientation scale scores of freshmen, sophomores, juniors, and seniors in the cross-validation group are presented in Table 1. Each class mean was progressively higher than that of the class below, ranging from a mean score of 132.9 for freshmen to a mean score of 193.5 for seniors. An analysis of variance indicated that the observed differences between the class mean scores were highly significant ($F = 34.7$, $df = 3/240$; $\alpha \leq .01$). Pairwise comparisons between mean scores of freshmen and sophomores, between sophomores and juniors, and between juniors and seniors, etc. were made, using Scheffe's test (Hays, 1963). These results show that the mean score of each class on the nurses' professional orientation scale was significantly greater ($\alpha \leq .05$) than the mean of the class below. A mean increase of 17.4 points was observed between the junior and senior classes; an increase of 21.8 points, between sophomores and juniors; and an increase of 21.4 points, between freshmen and sophomores. The greatest difference (60.6 points) was found between the senior and freshmen classes.

These results indicated that the nurses' professional orientation scale measures a shift in the student's view of the profession that was positively related to the length of the training experience. The congruence between student and faculty ratings of items of the nurses' professional orientation scale increased significantly with each year of training experience.

The reliability of this 60-item scale was computed using the responses of the cross-validation group. The internal consistency coefficient of the nurses' professional orientation scale, computed with the generalized coefficient alpha formula, was $r = .89$, indicated that the attitude measured by the nurses' professional orientation scale was a fairly unitary trait among student nurses.

DISCUSSION

Generally, the results of this study support the contentions of Davis and Oleson (1964) and Sharp and Anderson (1972) that there is a definite trend for the student to endorse the faculty view of the profession as graduation nears. This may indicate that the student's views are molded by actual clinical experiences, or that the faculty who teach students basic skills and procedures also indoctrinate them with their own professional views.

One limitation inherent in this study should be noted. All students in this study were enrolled in baccalaureate university programs. Although these programs are quite similar to many others throughout the country, the professional values adopted by students in other settings may differ. Particularly, these findings might not be applicable to students in junior college or hospital nursing programs.

The reliability and validity data for the nurses' professional orientation scale are very encouraging. Certainly, this instrument merits further experimental study. An interesting area for investigation is the relationship between clinical practice ratings and scores on the nurses' professional orientation scale. It would also be useful to learn the relationship between the student nurses' professional views and persistence in the profession. Additional studies such as these are needed before this scale could be used for individual counseling or evaluation.

Finally, it must be stressed that the score on the nurses' professional orientation scale measured only the degree of the student's conformity to the faculty's professional standards. The cause of a low score could not be inferred from the scale alone. A student could earn a low score by being too cynical about her profession, too naive about her profession, or too idealistic. Each potential user of this scale needs to examine the content of the items themselves to determine whether conformity to the professional views covered by these items is desirable in his own educational setting.

TABLE 1

CROSS-VALIDATION SAMPLE MEANS AND STANDARD DEVIATIONS ON NURSES PROFESSIONAL ORIENTATION SCALE SCORES FOR FOUR CLASSES OF STUDENT NURSES

Measure	Seniors	Juniors	Sophomores	Freshmen
$\bar{X}$	193.5	176.1	154.3	132.9
SD	30.3	33.1	30.3	22.8
Sample size	41	82	80	41

REFERENCES

DAVIS, F., & OLESEN, V. Baccalaureate students' images of nursing; a study of change, consensus, and consonance in the first year. *Nursing Research*, 1964, **13**, 8–15.

FOX, D. J., DIAMOND, L. K., & WALSH, R. C. *Satisfying and stressful situations in basic programs in nursing education*. New York: Institute of Research and Service in Nursing Education, Teachers College, Columbia University, 1960.

HAYS, W. L. *Statistics for psychologists*. Chicago: Holt, Rinehart & Winston, 1963.

KATZELL, M. E. Expectations and dropouts in schools of nursing. *Journal of Applied Psychology*, 1968, **52**, 154–157.

KIBRICK, A. K. Dropouts from schools of nursing: The effect of self and role perception. Unpublished doctoral dissertation, Harvard University, 1958.

OLESEN, V. L., & DAVIS, F. Baccalaureate students' images of nursing; a follow-up report. *Nursing Research*, 1966, **15**, 151–158.

POLLACK, S., & MICHAEL, W. B. Shifts in measures of attitudes of medical students toward those of their professors relative to the doctor image and the doctor-patient relationship: Implications for prediction of a clinically oriented criterion. *Educational and Psychological Measurement*, 1966, **26**, 1069–1072.

SHARP, W. H., & ANDERSON, J. C. Changes in nursing students' descriptions of the personality traits of the ideal nurse. *Measurement and Evaluation in Guidance*, 1972, **5**(2), 339–344.

SIEGEL, H. Professional socialization in two baccalaureate programs. *Nursing Research*, 1968, **17**, 403–407.

(Received December 11, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 2, 236–238

STRUCTURE AND MEASUREMENT OF MARINE RECRUIT ATTITUDES [1]

RICHARD F. BOOTH [2] AND ANNE HOIBERG

Navy Medical Neuropsychiatric Research Unit, San Diego, California

Five orthogonal factors labeled as attitudes toward the toughness of Marines, the spirit among Marines, affiliation with the Marine Corps, and authority and consideration in the Marine Corps were identified in the responses of recruits to items in a Marine Corps Opinion Questionnaire. Scales developed to provide scores on these factors should be useful in future studies of Marine-recruit attitudes and the relationship of these attitudes to posttraining effectiveness criteria.

Marine recruits come from a wide variety of backgrounds into the highly structured environment of basic training where, during the period of a few weeks, they are exposed to an intensive military indoctrination. An implicit objective of this indoctrination is to modify recruit attitudes, presumably for the purpose of improving subsequent performance and conduct in the Corps. Evidence, however, as to the nature of attitude changes that actually occur during this period and particularly the relationship of these attitudes and attitude changes to criteria of effective service, has generally been lacking. One reason for the lack of empirical evidence pertaining to these relationships has been the absence of a reliable measure of relevant recruit attitudes (Wilkins, 1969).

The purpose of this study was to develop a measure of attitudes toward the Marine Corps that could be used in a larger research project designed to explore Marine recruit attitudes, attitude changes, and the validity of these attitudes for predicting posttraining effectiveness. The primary objective of this larger research program was to evaluate these attitudes as a possible addition to the composite of variables currently employed for the prediction of effective service in the Marine Corps (Berry & Nelson, 1967; Plag, Goffman, & Phelan, 1971). Furthermore, it was expected that the identification of factors related to recruit attitudes and attitude changes eventually leads to improved decisions regarding programs designed to facilitate recruit adjustment to military life. The initial step in this overall investigation, therefore, involved the construction of an instrument that could be used to assess Marine recruit attitudes on several occasions during basic training.

One useful method for identifying the attitudes measured by this questionnaire and the items that constitute stable indices of these attitudes is factor analysis. The purpose of this study, then, was (*a*) to develop factor scales that provide stable indices of attitudes assessed consistently across administrations and (*b*) to evaluate the independence and reliability of scores computed on these factor scales.

METHOD

Subjects

The sample was composed of 481 Marine recruits, from 8 platoons, who entered training at the Marine Corps Recruit Depot (MCRD), San Diego, during October 1969 and who remained with the same platoon throughout training

Attitude Measure

A Marine Corps Opinion Questionnaire (MCOQ) was created to assess attitudes toward several aspects of Marine Corps life: (*a*) leadership, (*b*) discipline, (*c*) affiliation, (*d*) toughness, (*e*) group living, and (*f*) general living and working conditions. These areas were chosen, on the basis of interviews conducted and literature reviewed, because they were considered relevant to all Marines. After a sizable item pool had been assembled, the attitude items were evaluated for readability and meaningfulness by a number of Marine recruits and noncommissioned officers. Finally, 96 items were selected to be included in the experimental questionnaire. A six-category, Likert-style response format was employed with item alternatives ranging on a continuum from 1 (agree completely) to 6

---

[1] This study was supported by the Bureau of Medicine and Surgery, Department of the Navy, under Research Task MF12.524.001-9005. The opinions expressed are those of the authors and are not to be construed as necessarily reflecting the official policy of the naval service.

[2] Requests for reprints should be sent to Richard F. Booth, Navy Medical Neuropsychiatric Research Unit, Department of the Navy, San Diego, California 92152.

(disagree completely). To reduce the possible effects of response bias, half of the items selected were worded positively with respect to the Marine Corps, whereas the remaining items were worded negatively.

Procedure

The questionnaire was administered to the subjects four times during training; these administrations corresponded to the beginning of training, the end of the first and second phases, and the conclusion of training.

Administration of these questionnaires was conducted by a senior noncommissioned officer in a classroom at the Marine Corps Recruit Depot. It was emphasized in the instructions that answers would be treated as confidential and that no Marine would see the responses of an individual recruit.

RESULTS

The principal-axis factor model (Harman, 1967) was used to determine the factor pattern of item intercorrelations for each administration of the Marine Corps Opinion Questionnaire. The extraction of factors was stopped at 10 because, on the basis of limitations imposed on item content during creation of the questionnaire, it was not expected that additional factors would be useful for the purpose of this study. Further, since it was found that the percentages of variance explained by Factors 6 through 10 were very small, only the first five factors from each analysis were rotated to a normalized varimax criterion of simple structure (Kaiser, 1958). Coefficients of congruence computed to determine the degree of similarity between rotated factors from each analysis (Harman, 1967) suggested that the same five factors were represented in the attitudinal responses made by recruits during each administration.

The five factors extracted from the Marine Corps Opinion Questionnaire responses of recruits may be labeled as attitudes toward toughness, spirit, affiliation, authority, and consideration. Items correlated with Factor 1 (9%),[3] toughness, pertain to the predisposition of a Marine toward hostilities as well as the invincibility of a Marine once in combat. Factor 2 (11%), spirit, was loaded with items which suggested a cynical view of Marines as people. When the wording of these items was reversed from negative to positive, however, they seemed to suggest a spirit that may exist among Marines, that is, that Marines are honest, friendly, loyal to their buddies, hardworking, ambitious, and brave. Items with high loadings on Factor 3 (9%), affiliation, refer to serving in, and identifying with, the Marine Corps. The items loading on Factor 4 (9%), authority, pertain to traditional views of leadership and discipline in the Marine Corps. Finally, the content of items that were correlated with Factor 5 (8%), consideration, reflect a concern on the part of leaders for the needs and problems of subordinate Marines.

In order to develop scales that could be used to measure these factors, items were selected from the Marine Corps Opinion Questionnaire which had consistently high loadings on one of the factors and low loadings on the others. Some compromises were made, however, when an item-loading was low on a factor on a single administration or when an item had a high loading on more than one factor, as occurred in several instances with Factors 4 and 5 on the fourth administration. In this manner, 35 items were chosen as indexes for the five factors; representative items from each scale with corresponding fourth administration factor loadings are presented below.

Representative Items and Factor Loadings
For Each of the Five Attitude Scales

Factor 1 (7 items): Toughness
2. Combat is not hazardous for a well-trained Marine. (.53)
45. Most Marines would eagerly take on several civilians at a time in a bar fight. (.62)

Factor 2 (10 items): Spirit
13. In the Marine Corps, nobody seems to "give a damn" about anything. (.57)
85. A smart Marine does only what is required for him to "get by." (.64)

Factor 3 (5 items): Affiliation
58. If I had my choice now I would prefer to be a civilian. (.88)
73. Despite everything, I feel more at home in civilian life. (.72)

Factor 4 (10 items): Authority
20. The most important thing to teach Marines is absolute obedience to their superiors. (.52)
68. Any good Marine Corps officer or NCO should be strict with the men under him in order to gain their respect. (.64)

Factor 5 (3 items): Consideration
91. In general, Marine Corps officers and NCOs keep their men informed about what is going on. (.65)
95. Marine Corps officers and NCOs are usually understanding of the needs and problems of their men. (.68)

---

[3] Percentage of variance is explained by this factor on fourth administration of the Marine Corps Opinion Questionnaire.

The final objective of this study was to determine the intercorrelations and reliabilities of scores computed as indices of these factors. Factor scores were computed for each administration by reflecting (subtracting from 7) responses to items worded negatively with respect to the Marine Corps and then summing responses across all items chosen to represent a factor. Intercorrelations among these scores were computed and reliability coefficients were derived for each scale. These figures indicated that the intercorrelations among scores were low to moderate in magnitude (average $r = .34$) and that the estimates of test-retest reliability computed for each scale were quite high (average reliability coefficient = .82).

DISCUSSION

The consideration and authority factors identified in the results of this study were strikingly similar to the "consideration" and "initiating structure" dimensions of leader behavior which emerged from the Ohio State Leadership Studies (Stogdill & Coons, 1957). The "consideration" factor pertains to the amount of friendship, mutual trust, respect, and warmth generated between a leader and his group; the "initiating structure" factor pertains to the degree of definition by a leader of group organization, group activities, and of his relationship to the group. Although the Ohio State factors have been identified in a wide variety of leadership situations (Fleishman, 1973; Stogdill & Coons, 1957), the present study suggests that the generality of these two constructs may extend to a highly structured, authority-oriented environment such as Marine Corps basic training. Of more importance, however, is the finding in several studies that the amount of "consideration" displayed by leaders is inversely related to such behavioral measures as grievances, absenteeism, and turnover (see e.g., Fleishman & Harris, 1962; Fleishman, Harris, & Burtt, 1955; Vroom, 1964). These relationships suggest that investigating the potential uses of consideration scores for evaluating and predicting leader effectiveness in the Marine Corps may be a promising area for further research.

Scale reliabilities were high, despite the relatively small number of items representing each factor. Nevertheless, additional items should be developed to supplement each scale. Furthermore, a better balance is needed in each scale between positively and negatively worded items.

Overall, the findings of this study have demonstrated that the Marine Corps Opinion Questionnaire can indeed be a useful measure of attitudes toward the Marine Corps. The attitudes assessed by this questionnaire appear to be relevant with respect to the Corps and the scales developed to measure these attitudes were found to be reliable. The Marine Corps Opinion Questionnaire, then, should be particularly useful for the study of Marine recruit attitudes. Furthermore, scores derived from this measure should be evaluated as potential predictors of effectiveness criteria.

REFERENCES

BERRY, N. H., & NELSON, P. D. Combat adjustment and four-year effectiveness of Marine Corps enlistees. (Report 67-17) San Diego, Calif.: Navy Medical Neuropsychiatric Research Unit, 1967.

FLEISHMAN, E. A. Twenty years of consideration and structure. In E. A. Fleishman & J. G. Hunt, *Current developments in the study of leadership.* Carbondale, Ill.: Southern Illinois University Press, 1973.

FLEISHMAN, E. A., & HARRIS, E. F. Patterns of leadership behavior related to employee grievances and turnover. *Personnel Psychology,* 1962, **15**, 43–56.

FLEISHMAN, E. A., HARRIS, E. F., & BURTT, H. E. *Leadership and supervision in industry.* Columbus: Ohio State University, Bureau of Educational Research, 1955.

HARMAN, H. H. *Modern factor analysis.* Chicago: University of Chicago Press, 1967.

KAISER, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika,* 1958, 23, 187–200.

PLAG, J. A., GOFFMAN, J. M., & PHELAN, J. D. Predicting the effectiveness of new mental standards enlistees in the Marine Corps. (Report No. 71-42) San Diego, Calif.: Navy Medical Neuropsychiatric Research Unit, 1971.

STOGDILL, R. M., & COONS, A. E. (Eds.) *Leader behavior: Its description and measurement.* Columbus: Ohio State University, Bureau of Business Research, 1957.

VROOM, V. H. *Work and motivation.* New York: Wiley, 1964.

WILKINS, W. L. Attitudes and values as predictors of military performance. In N. A. B. Wilson (Ed.), *Manpower research.* London: English Universities Press, 1969.

(Received February 12, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 2, 239–240

A DISGUISED MEASURE OF CIVIL RIGHTS ATTITUDES

ABRAHAM K. KORMAN[1]

Baruch College, City University of New York

The purpose of the research was to construct and to validate two knowledge tests at different occupational levels as disguised measures of civil rights attitudes. Results indicated that there is a significant relationship between knowledge of civil rights events, attitudes toward civil rights, and membership in groups concerned with civil rights affairs. These findings seem sufficiently promising to justify further research.

The development of a measure of civil rights attitudes that is free of "social-desirability" seems to be of value for two reasons. First, it should provide some guidance toward resolving the persisting problem of how psychologists can effectively measure individual differences along significant emotional and attitudinal dimensions. Second, it should be a step toward helping administrators identify for appropriate action those individuals who might be antagonistic toward the full integration of our nation's organizations.

The approach that was decided on to develop such a measure was based on the logic used by Greene (1940) in his development of the Michigan Vocabulary Profile Test. This test is based on the logic that we become more knowledgeable in those fields where we have positive interest, and evidence has been presented which supports this assumption (Greene, 1940). More recently, Ewen (1969) has used a similar logic in explaining his obtained positive correlation between scores on the Graduate Record Examinations (GRE) Advanced Tests: Psychology and performance in graduate school in psychology, *after* verbal intelligence had been partialled out.

These studies served as a background for development of a series of test items that measured knowledge of civil rights events and the ability to think logically with such materials. The hypothesis tested was that scores on such measures are related to independent measures of civil rights attitudes and behaviors.

METHOD

Test measures

After pretesting, two separate tests were developed, a long difficult 69-item form and a short, easier 29-item form. Both versions were designed to reflect the subject's knowledge of civil rights events and his logical thinking adequacy when presented with civil rights materials.

Samples and Administration

The tests were administered to various groups. For student groups, administration took place in class sessions and as part of research requirements for introductory psychology. For industrial groups, volunteers were obtained with company cooperation with responses by mail. All tests were administered anonymously.

While in no way could the groups be considered random samples from particular populations, the following fairly diverse groups of respondents were obtained: (*a*) 87 financial-district managers, all of whom were enrolled in the evening master's of business administration program of a local university; (*b*) 24 managers in the insurance industry; (*c*) 52 commerce students at a suburban university under religious control; (*d*) 156 liberal arts (mostly psychology majors) students at two urban universities; this group included a subsample of 73 within which a sufficient number of individuals belonged to civil rights groups to enable the use of this variable as a second criterion measure. In addition, Wonderlic scores were available to be used as control measures in another group of 44 from this overall sample of 156; (*e*) 25 lower level insurance company employees; (*f*) 28 employees of diverse occupational levels from an office supplies corporation; and (*g*) 140 evening students at a local community college, almost all of whom were employed in subprofessional lower level positions.

Samples *a*, *b*, *c*, and *d* responded to the long form while *e*, *f*, and *g* responded to the short form.

Civil Rights Indexes

At the close of each testing session or at the end of the test, each respondent was asked to indicate his answer to two questions: (*a*) Did he belong to any group that could be considered a

---

[1] Requests for reprints should be sent to Abraham K. Korman, Department of Psychology, Baruch College, City University of New York, 17 Lexington Avenue, New York, New York 10010.

TABLE 1

RELATIONSHIPS BETWEEN CIVIL RIGHTS KNOWLEDGE AND ATTITUDES FOR DIFFERENT GROUPS

Sample	Relationship			
	n	$\bar{X}$	*SD*	*r* with civil rights attitudes
Long test				
Managers	87	33.4	10.8	.27**
Insurance managers	24	36.7	8.4	.40*
Suburban university	52	31.8	8.8	.42**
Urban university	156	31.6	7.1	.33**
Urban university	73	35.0	6.9	.27*[a]
Short test				
Insurance company employees	25	23.0	3.8	.40*
Office equipment company employees	28	22.7	3.7	.33
Mixed employee sample	140	17.2	4.1	.46**

[a] *r* biserial with "membership in civil rights groups."
* $p < .05$.
** $p < .01$.

pro-civil rights group? (*b*) How would he rate himself on a 5-point scale as to his attitude toward civil rights (1 = anti, 5 = pro, 3 = neutral, etc.)? This latter question, with the one exception previously mentioned, was the major dependent variable in this research.

RESULTS

The results of the research are outlined in Table 1. In general, these findings justify the conclusion that there is a significant relationship between knowledge of civil rights events, attitudes towards civil rights and membership in groups concerned with civil rights affairs. This conclusion did not appear limited by the nonlinearity of regressions, since scatter plot analyses indicated that in no case were the regressions nonlinear; nor did it appear to be limited by the contaminating influence of intelligence as the major contributor to the correlation between the knowledge test and the attitude measurement. Thus, for a subsample of 44 within the urban university sample of 156, the correlation between the knowledge test and the intelligence test was not significant ($r = .29$, $p > .05$). For the same group, the correlation between the Wonderlic and the attitude measurement was +.00, and the partial correlation coefficient, between the knowledge test and attitudes, with the Wonderlic scores partialled out, was +.33 ($p < .05$).

DISCUSSION

It would seem that the implications of Greene's (1940) and Ewen's (1969) work for developing measures of attitudes received support in this study. However, further research does seem warranted as to why the relationships, while consistent, are sometimes modest in magnitude. One explanation might be the possible unreliability of our criteria (i.e., 1-item attitude measurement and statements of membership). Perhaps stronger criteria may have led to stronger relationships.

Even with stronger criteria, however, it is likely that the practical value of an instrument of this nature is limited to samples relatively homogeneous in education. The reason for this limited use is the relatively well-substantiated finding of a general relationship in the population between educational level, liberal racial attitudes and general knowledge (cf. Allport, 1958, pp. 405–407). Yet, this finding does not obviate the fact that, holding educational level constant, there may be other variables, such as positive interest, that may affect the relationship between knowledge and attitudes among individuals within each educational level. Such positive interest could lead to the types of relationships which we have reported here, relationships which, while encouraging, still need to be further validated with behavioral criteria.

REFERENCES

ALLPORT, G. W. *The nature of prejudice.* Garden City, New York: Doubleday, 1958.

EWEN, R. The GRE Psychology Test as an unobstrusive measure of motivation. *Journal of Applied Psychology,* 1969, **53**, 383–387.

GREENE, E. B. Vocabulary profiles of groups in training. *Journal of Educational Research,* 1940, 33, 569–575.

(Received December 4, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 2, 241–243

SELF-IMAGE AND STEREOTYPES OF FEMININITY: THEIR RELATIONSHIP TO WOMEN'S ROLE CONFLICTS AND COPING[1]

FRANCINE E. GORDON[2]
Stanford University

DOUGLAS T. HALL
Michigan State University

The relationships of self-image, the female image of a feminine woman, and the female perception of the man's image of a feminine woman to role conflicts, coping behavior, satisfaction, and happiness in a sample of 229 college-educated women were explored and were found to be related. The woman's perception of the male's stereotype of femininity was the best predictor of various types of conflicts experienced. The woman's style of coping with conflict was a function of her self-image. Self-image was also associated with satisfaction and happiness. The impact of male stereotypes on women's conflicts was discussed in terms of current social definitions of women's roles.

The consciousness-raising activities of women's liberationists have resulted in a renewed interest in women's roles, the conflicts inherent in those roles, and the behavior which copes with such conflicts. In an earlier study of conflicts in women, one of the present authors classified women's conflicts by sources: self, home roles, and nonhome roles (Hall, 1970). Conflict may arise from pressures between different home roles, between different nonhome roles, between home and nonhome roles, between different needs of the self, or between the self and home or nonhome roles.

Hall & Lawler (1971) also classified methods of coping with role conflicts as follows: Type 1, structural role redefinition (redefinition of expectations held by others); Type 2, personal role redefinition (modification of one's own expectations of one's own behavior); Type 3, responsive role performance (attempting to do everything expected without changing others' expectations).

The present study explored the relationships of types of role conflicts and methods of coping behavior to (*a*) the woman's self-image, (*b*) her image of a feminine woman, and (*c*) her perception of the male image of a feminine woman[3]. These two variables, conflict type and coping method, were also related to the discrepancy of a feminine woman or her understanding of the male image of a feminine woman. Additionally, the relationships of these three image variables with satisfaction and happiness were investigated.

METHOD

Instrument[4]

A questionnaire was developed for use in this study. The first section asked for demographic information to divide the sample according to marital status. The second part of the questionnaire assessed stereotypes and self-image using three sets of semantic differentials. The first semantic differential asked the woman to rate herself on 28 bipolar adjective scales. Each scale had seven positions (1 to 7 points). The second semantic differential asked the subject to rate her image of a feminine woman on 27 adjective scales (excluding the pair "feminine–unfeminine" because the woman was rating a "feminine" woman). Finally, each subject was asked to rate what she perceived to be the average man's image of a feminine woman, using the same 27 pairs of adjectives which she had used to rate her own image of a feminine woman.

Each semantic differential was factor analyzed separately. The same four factors emerged in each case (Gordon, 1971a). The factors were as follows: (*a*) potency: emotional and physical strength; (*b*) supportiveness: nurturant, agreeable behavior in interpersonal settings; (*c*) spontaneity: willingness to take immediate action; and (*d*) emotionality: level of emotion.

The third section consisted of the following questions: (*a*) What conflicts or strains, if any, have you experienced between your various roles in life? (*b*) How do you deal with these conbetween her self-image and either her own image flicts? (*c*) Overall, how satisfied do you feel with the way you deal with your roles in life? (*d*) In general, how happy would you say you are?

---

[1] This research was supported in part by a Ford Foundation grant to the Department of Administrative Sciences, Yale University. The cooperation of the Alumni office of the University of Connecticut is gratefully acknowledged.

[2] Requests for reprints should be sent to: Francine E. Gordon, Graduate School of Business, Stanford University, Stanford, California, 94305.

[3] Hawley (1971) has shown that role behavior of women is strongly affected by their perception of the male's stereotype of women, regardless of the actual stereotypes men have.

[4] Details about the instrument, scoring or analysis will be made available upon request.

Sample

Ninety female graduates of the liberal arts college at the University of Connecticut were selected randomly from each of five graduating classes. Analyses showed no significant differences between classes, and data presented in this study include members of all classes.

Of the 450 forms mailed out, 261 (58%) were completed and returned. Of these, 56 were from the class of 1948, 49 from 1953, 55 from 1958, 54 from 1963, and 47 from the class of 1968. No demographic information was available from the university to determine if the sample was representative. Among the 261 returns, 229 had been completed by women who were married. These women constituted the present sample.

Scoring and Analysis

Four scale scores corresponding to the four factors were computed for each person on each semantic differential. Each scale consisted of adjective pairs that had their highest loading on the corresponding factor. Then, the score for each factor scale on the self-image rating was subtracted from the corresponding scale score on the rating of the woman's own image of a feminine woman. In a similar fashion, each scale score on the self-image rating was subtracted from the corresponding factor score on the rating of woman's perception of the male image of a feminine woman.

Responses to the question on conflict (section 3, question 1) were coded using the Hall & Lawler (1971) classification system described earlier. If the subject indicated that she did not experience conflict, she was given a score of 1 for the category "no conflict"; if she reported conflict, she was given a score of 0. Each conflict was classified in terms of the competing pressures which gave rise to it. For example, a conflict between the roles of wife and mother should be put in the home versus home category because it involved pressures from two roles within the home. The remaining five categories were home versus nonhome, home versus self, nonhome versus nonhome, self versus nonhome, and self versus self (intrapsychic). Conflicts were scored dichotomously with the subject given 1 if she experienced a particular type and 0 if she did not.

The individual pressures leading to conflict were also coded for each respondent, indicating whether or not she had any conflicts involving home pressure, nonhome pressures, and/or pressure from the self.

Methods of coping were classified into three types according to the Hall (1970) scheme described earlier. The coding scores for each woman were the number of each type of coping that she employed. Sixteen strategies for coping with role conflicts were identified. On samples of 15 and 30 questionnaires, two coders agreed 61% and 74%, respectively, on responses. Next, the 16 coping strategies were allocated to the three types of coping referred to earlier so that each woman received one score for the number of Type 1 coping strategies, one score for Type 2 strategies, and one score for Type 3 strategies.

The question on satisfaction was scored using a scale from 1 (dissatisfied) to 5 (extremely satisfied). Happiness was also scored from 1 (very unhappy) to 5 (very happy). The use of general single item questions to measure these two global constructs is based on the work of Bradburn (1969) and Gurin, Veroff, and Field (1960).

RESULTS[5]

Self-Image and Stereotypes Versus Conflicts

Almost one fourth of the sample (24%) reported no conflicts. The most common conflict was that between a home role and a nonhome role (31%). The most common source of pressure was the woman's home roles (64%).

None of the self-image or stereotype scores correlated significantly with the existence or nonexistence of conflict per se. The woman's image of the male's image of a feminine woman was the one most closely associated with conflicts experienced by the woman. All four factor scales of the perceived male stereotype correlated significantly with home versus nonhome conflicts ($-.20$, $p < .01$; $-.16$, $p < .05$; $-.15$, $p < .05$; $.18$, $p < .01$) and with nonhome only pressures ($-.19$, $p < .01$; $-.15$, $p < .05$; $-.16$, $p < .05$; $.18$, $p < .01$). Supportiveness also was related to self versus home conflict ($r = .18$, $p <$ 01) and self-only pressures ($r = .20$, $p < .01$). Thus, the less potent and more emotional a woman believed men felt women were, the more conflicts she experienced in the home versus nonhome category.

Correlations between the discrepancy scores between self-image and perceived male image of a feminine woman and conflicts were similar in pattern to, but lower than, those involving the perceived male image by itself.

---

[5] Complete tables will be made available upon request.

The discrepancy between the woman's self-image and her image of a feminine woman was more strongly related to conflicts than was either self-image or her image of a feminine woman. Supportiveness seems to be an important dimension in those discrepancies. The less supportive a woman felt relative to her stereotype of femininity (greater the discrepancy), the less likely she was to experience nonhome pressures (−.19, $p < .01$), and the more likely she was to experience home versus home conflict (.16, $p < .05$).

Self-Image and Stereotypes Versus Coping

The best predictor of coping behavior in this study was the woman's image of herself and in particular her perceived supportiveness. There was a tendency for a woman who perceived herself as more supportive to cope with conflict by structurally redefining her roles (Type 1, $r = .16$, $p < .05$), and not attempting to meet all demands placed on her (Type 3, $r = -.20$, $p < .01$).

Self-Image and Stereotypes Versus Outcome

The best predictor of both satisfaction and happiness was the woman's image of herself (which was a poor predictor of conflicts). The more potent, supportive, and unemotional the woman felt, the more satisfied ($r = .32$, .30, −.24, respectively, $p < .01$), and the happier she was ($r = .25$, .20, −.25, respectively, $p < .01$). The woman's perceived male image of a feminine woman and her own image of a feminine woman were not correlated with her satisfaction and happiness.

DISCUSSION

Of all the various measures and discrepancy scores used in the present study, the woman's perception of the male stereotype was overall the most closely associated with conflicts. In particular, the number of conflicts involving nonhome roles tended to decrease the more potent, supportive and unemotional the woman perceived the male's stereotype of femininity to be. This finding suggests that in the more male-dominated environment outside the home, the man's standards of femininity may create difficulties for women. It is interesting that women who work (i.e., have more nonhome involvement) perceive a greater difference between their own image of a feminine woman and that of the males than do housewives (Gordon, 1971b). The former attribute men with a more conventional sex role stereotype of woman, for example, low in potency and high in emotionality. Whether accurate or not, this perception could create psychological pressures for the working woman as she tries to be both competent and feminine.

With regard to self-related conflicts, the perceived male stereotype of femininity had relatively the strongest influence of the variables measured. This finding lends support to the arguments of women activists that men control women not only through discriminating behavior, but also through their impact on women's attitudes and inner conflicts.

Whatever male influence exists appears to have its limits. Ultimately, it seems to be the woman's own self-image that determines her overall satisfaction and happiness.

REFERENCES

BRADBURN, N. *The structure of psychological well-being*. Chicago, Ill.: Aldine, 1969.

GORDON, F. E. What women are made of: A factor analytic study. Unpublished paper, Department of Administrative Sciences, Yale University, 1971 (a).

GORDON, F. E. Where happiness lies: Self-image, stereotypes, work-status, and happiness. Working paper, Department of Administrative Sciences, Yale University, 1971 (b).

GURIN, G., VEROFF, J., & FIELD, S. *Americans view their mental health*. New York: Basic Books, 1960.

HALL, D. T. Role conflict, coping and satisfaction through the life cycle: The careers of women. Unpublished paper, Yale University, 1970.

HALL, D. T., & LAWLER, E. E. Job pressures and research performance. *American Scientist*, 1971, 59, 64–73.

HAWLEY, P. What women think men think. *Journal of Counseling Psychology*, 1971, **3**, 193–199.

(Received January 2, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 2, 244-246

EFFECTS OF EXPECTATION, SOCIAL CLASS, AND EXPERIENCE ON IN-SERVICE TEACHER BEHAVIOR IN SMALL GROUPS[1]

LEN SPERRY[2]

Marquette University

Eighty-one teachers were randomly assigned to planning groups of three and then randomly assigned to one of four treatments to test the hypotheses that expectation and social class affected small group behavior. Two levels of expectation and two levels of school social class were simulated. Results indicated that the instructional material produced by groups exposed to the high-expectation treatment rated significantly higher ($p < .01$) than that of groups exposed to the low level; and that groups exposed to the high-expectation, upper-social-class treatment produced instructional materials that were more intellectual than vocational in orientation.

A considerable amount of research has been reported concerning variables that affect the instructional planning done by in-service teachers. Often, such research is both atheoretical and noncomprehensive, since it has focused on only one set of variables: role, cultural, or personal variables. Subsequently, such studies offer only partial explanations and a limited basis for generalizations and subsequent research. Thus, there is ample indication that a theoretical and comprehensive framework is required for a better understanding of in-service teacher behavior.

Getzels (1963, 1968, 1969) has offered a general yet comprehensive sociopsychological model for understanding social behavior in institutional settings. Basically, Getzels's model assumes that human behavior is determined by both normative, or sociological, and personal, or psychological, factors and that these factors arise out of a cultural, or anthropological, context. Within an institution, such as a school, there are expectations shared for various roles, such as teacher, administrator, and student, and for personal need dispositions. The framework, based on the sociological theory of Parsons, suggests that both the institution and the individual are embedded in culture or subcultures, each with its own values and ethos. These factors interact in the determination of social behavior.

More than 100 educational studies have been based on Getzels's social systems model (Getzels, Lipham, & Campbell, 1968; Wahlberg, 1969). The present investigation was likewise based on Getzel's model and studied three sets of variables that might influence teacher in-service behavior. The effects of the following variables on in-service teacher behavior were examined: (*a*) a cultural variable—school social class level, (*b*) a role variable—administrative expectation, and (*c*) a personal variable—years of experience in teacher planning.

METHOD

Design

To determine whether administrative expectations could have a significant effect on an assigned instructional planning task, the effects of two treatments of administrative expectation on the outputs of a group task were observed. The two treatments were designated as high- and low-administrative expectation. The first major hypothesis suggested that high expectation elicits higher output than low expectation. The actual writing of an instructional unit by each group constituted the group instructional task, which was rated in terms of a quantified index, the Rating Scale for Curriculum Evaluation (Talmage, 1970) and yielded a total output, the group instructional task score.

A secondary consideration of this study related to the social class level of the school and the community it served. The second major hypothesis suggested that teachers in upper-class schools produce higher group instructional task scores than teachers in lower-class schools. Upper was designated to mean upper-middle and upper class, while lower was designated to mean lower and lower-middle class. An upper- and a lower-class school were simulated to ascertain whether social class might significantly affect group instructional task scores under the two treatments of administrative expectation. Furthermore, a corollary hypothesis suggested that teachers in upper-class schools produce an instructional unit that is primarily intellectually oriented, whereas teachers in lower-class schools produce an instructional unit that is primarily vocationally oriented. The selection of variables yielded a 2 × 2 covariate design, with two levels of administrative expectation and two levels of social class, years of experience being the covariate.

---

[1] This article was based on a doctoral dissertation submitted to the Graduate School, Northwestern University, June 1970. An earlier version of this article was presented at the American Educational Research Association meeting, February 1971.

[2] Requests for reprints should be sent to Len Sperry, Department of Educational Psychology, Marquette University, 502 North 15th Street, Milwaukee, Wisconsin 53233.

Simulation Materials

One set of simulation materials was produced for each of the four treatments. These materials were similar in format to those included in the In-Basket Test devised by Hemphill, Griffiths, and Frederiksen (1962). The simulated materials included film strips, newspaper clippings, minutes from faculty meetings, and pertinent demographic data. In addition, audiotaped instructions served to standardize the treatment conditions in terms of timing, voice, and emphasis. These directions, along with the other simulation materials, conveyed the net impression that in-service instructional planning was of primary concern, or of little concern to the administrator of the particular school, in terms of the upper- or lower-class setting of the administrator's school.

Source of the Data

Data for this study were obtained from two sources: (*a*) from each participant's Personal Inventory Form which indicated the following eight variables: age, sex, education level, teaching experience, curriculum experience, subject teaching experience, grade level experience, administrative experience; and (*b*) from solutions to the group instructional task and quantified by the Rating Scale for Curriculum Evaluation.

Reliability for the Rating scale, determined by Hoyt's estimate of test reliability based on the analysis of variance (Holt & Stunkard, 1952), was .94. Rater reliability among the three raters who evaluated the data was .88 ($p < .01$).

Sample

Eighty-one volunteer and nonvolunteer teachers were randomly assigned to groups of 3. Nineteen of the participants were members of minority groups, of which 16 were blacks. Chi-square tests to determine equalization of distribution of the eight characteristics on the Participant Inventory Form indicated no significant difference at the .05 level of significance. An F test at .01 level showed no significant differences between scores obtained by volunteer and nonvolunteers.

These 27 groups were then randomly assigned to the four treatments. In a one and one half-hour session, each group was exposed to the simulation and completed an assigned task that consisted of an upper grades, general social studies unit.

Other Preliminary Analysis of Data

The two assumptions underlying the use of the analysis of variance are normality of distribution and homogeneity of variance. Results from a probit analysis justified the assumption that data of this study represented an unbiased sample from a normal population. Results from Hartley's F_{max} justified the assumption of homogeneity of cell variance.

RESULTS

The main variables of this study were level of expectation and level of social class, while the suggested covariate was number of years of experience in instructional planning for each individual. In order to justify the use of a covariance design it had to be established that a sizable correlation existed between the covariate and the dependent variable. Since this correlation was quite low, the covariance design was replaced by 2 × 2 analysis of variance design that used the least squares method for unequal cell frequencies (Winer, 1962). Post hoc comparisons of cell means were performed.

First, it was found that groups exposed to the high level of administrative expectation rated significantly higher in terms of group instructional task scores than groups exposed to the low level of administrative expectation ($F = 37.04$, $df = 1, 77$, $p < .01$). Therefore, the first null hypothesis was rejected.

Second, it was found that groups exposed to the upper level of the school social class treatment did not rate significantly higher than groups exposed to the lower level of school social class ($F = 1.00$, $df = 1, 77$, $p < .01$). Therefore, the second null hypothesis was not rejected.

Third, it was found that groups exposed to the high-expectation, upper-social-class treatments significantly favored an instructional unit that was intellectually oriented over one that was vocationally oriented ($F = 9.70$, $df = 1, 77$, $p < .01$). Therefore, the third null hypothesis was rejected.

DISCUSSION

The theoretical significance of this study was its validation of a systems model that accounted for *one* specific variable in the personal, group, and cultural dimensional categories that are purportedly contingencies of group instructional planning of in-service teachers. Within the proposed social instructional systems model, further investigation could conceptualize and control for other variables suggested in terms of the personal, group, institutional, and cultural dimensional categories, thus promoting the kind of feedback necessary to further develop instructional theory and practice. The practical significance of this study was its demonstration that the building level administrator, the principal, is the real gatekeeper and quality control for educational change, irregardless of teacher experience and social class factors. Also, years of experience appear to be too undifferentiated a variable to have been meaningful in this study. Finally, the application of the noncomputer simulation technique developed for the present study suggested that such techniques offer the possibility not only of increasing the scope of theory but also of modifying in-service, as well as preservice, education by the use of such techniques for training and evaluation.

REFERENCES

GETZELS, J. Conflict and role behavior in the educational setting. In W. Charters & N. Gage (Eds.), *Readings in the social psychology of education.* Boston: Allyn & Bacon, 1963.

GETZELS, J., LIPHAM, J., & CAMPBELL, R. *Educational administration as a social science.* New York: Harper & Row, 1968.

GETZELS, J. A social psychology of education. In G. Lindzey & M. Aronson (Eds.), *The handbook of social psychology.* Reading, Mass.: Addison-Wesley, 1969.

HEMPHILL, J., GRIFFITHS, D., & FREDERIKSEN, N. *Administrative performance and personality.* New York: Columbia University Press, 1962.

HOYT, C., & STUNKARD, C. Estimating of test reliability for unrestricted Item Scoring Methods. *Educational and Psychological Measurement,* 1952, **12**, 756–758.

TALMAGE, H. Effects of verbal behavior within curriculum development committees on the curriculum product. Paper presented at the meeting of the American Educational Research Association, Minneapolis, March 1970.

WAHLBERG, H. Social environment as a mediator of classroom learning. *Journal of Educational Psychology,* 1969, **60**, 443–448.

WINER, B. *Statistical principles in experimental design.* New York: McGraw-Hill, 1962.

(Received December 18, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 2, 247–248

EFFECT OF TERMINAL RANKING ERROR ON MEANINGFUL OCCUPATIONAL CHOICE

EDWIN E. WAGNER[1] AND THOMAS O. HOOVER

University of Akron

The existence of terminal error (the tendency to overevaluate the first and the last positions in ranking a set of entities) was investigated in the ranking of occupational preferences for trustworthiness. It was found that terminal error occurred, but that it disappeared as intrinsic differences among the occupations to be ranked became more pronounced.

It has been shown that there is a tendency to overevaluate the first and last positions in ranking a set of five entities (Wagner & Hoover, 1974). However, the initial series of experiments illustrating this phenomenon dealt with words which were ranked according to "imagery value" and it seemed questionable whether these results could be extrapolated to more meaningful situations such as a foreman ranking his workers on productivity or a potential voter ranking a group of candidates. The present study explored the biasing effect of "terminal error" by asking subjects to rank occupations along a dimension of "trust"—a task which presumably evokes commitment and emotional involvement and might therefore more closely resemble real life situations where choices have practical consequences.

METHOD

In a study by Rotter and Stein (1971) subjects rated 20 occupations on the degree to which they felt that typical members of each occupational group were trustworthy and then rank-ordered the occupations from 1 through 20 on the degree of trust attributed to them. In the present experiment, the Rotter and Stein list of 20 occupations was subdivided into three separate lists of 5 occupations each and presented to three groups of 120 undergraduates from the University of Akron. Each list of 5 occupations was presented in completely counterbalanced order (in every possible combination, $5! = 120$ arrangements) with instructions to rate the occupations on trust. The three groups of subjects were instructed to give the highest trust rating the rank of 5; the occupation given the next highest trust rating was ranked 4, etc.; the rank of 1 was assigned to the occupation which was trusted the least. List 1 consisted of the first 5 occupations included in the original 20, that is, those ranked 20, 19, 18, 17, and 16 along the trust continuum: physicians, clergymen, dentists, judges, and psychologists. List 2 included those occupations rated 20, 17, 14, 11, and 8: physicians, judges, psychiatrists, law enforcement officials, and executives of large corporations. List 3 was comprised of those occupations rated 20, 16, 12, 8, and 4: physicians, psychologists, lawyers, executives of large corporations, and auto repairmen. By spacing the rankings in this manner, it was possible to study the influence of the degree of known differences among occupations on subsequent ranking error.

RESULTS AND DISCUSSION

Friedman two-way analyses of variance were computed to test the null hypothesis that ranking of the occupations would be unaffected by serial position. Table 1 reports the results along with the median value for each ranking position. The null hypothesis was rejected in List 1 beyond the .001 level but was accepted for Lists 2 and 3. Inspection of the medians in List 1 confirms previous research, inasmuch as the terminal points (ranks 1 and 5) were overranked in relation to the center positions. As might be anticipated, the greater the intrinsic differences among data, the less susceptibility to ranking error; Lists 2 and 3, which were more heterogeneous in original selection value than List 1, were unaffected by terminal error. However, much of the ranking tasks which are encountered in actual industrial and social situations (e.g., workers ranking supervisors or enlisted men ranking officers) probably involve fairly close discriminations where terminal error can be expected to occur. Since correction procedures for ranking error that have been presented in the literature (Rosander, 1936; Symonds, 1936; Thurstone, 1931) do not deal with terminal error, it seems that more attention should be

---

[1] Requests for reprints should be sent to Edwin E. Wagner, Department of Psychology, University of Akron, Akron, Ohio 44304.

TABLE 1

MEDIANS AND FRIEDMAN TWO-WAY ANALYSES OF VARIANCES COMPARING EFFECTS OF POSITION ON RANKINGS OF THREE SETS OF FIVE OCCUPATIONS POSSESSING DIFFERENT INTRINSIC RANGES OF TRUST VALUE

Experiment	Range of values	N	Ranking position					df	χ_r^2
			1	2	3	4	5		
1	16–20	120	3.14	2.83	2.92	2.94	3.21	4	71.4*
2	8–20	120	3.18	2.77	2.85	3.28	3.02	4	1.1
3	4–20	120	2.97	3.00	3.15	3.01	2.91	4	.5

* $p < .001$.

directed toward an analysis and rectification of this phenomenon.

REFERENCES

ROSANDER, A. C. The standard error of a mean rank order. *Journal of Educational Psychology*, 1936, **27**, 193–196.

ROTTER, J. B., & STEIN, D. K. Public attitudes toward the trustworthiness, competence, and altruism of twenty selected occupations. *Journal of Applied Social Psychology*, 1971, **1**, 334–343.

SYMONDS, P. M. Influence of order of presentation of items in ranking. *Journal of Educational Psychology*, 1936, **27**, 445–449.

THURSTONE, L. L. Rank order as a psychophysical method. *Journal of Experimental Psychology*, 1931, **14**, 187–201.

WAGNER, E. E., & HOOVER, T. O. The effect of serial position on ranking error. *Educational and Psychological Measurement*, 1974, in press.

(Received December 26, 1972)

Journal of Applied Psychology
1974, Vol. 59, No. 2, 249–252

DEVELOPMENT OF BEHAVIORALLY ANCHORED RATING SCALES AS A FUNCTION OF ORGANIZATIONAL LEVEL

SHELDON ZEDECK[1]
University of California, Berkeley

NICHOLAS IMPARATO
University of San Francisco

MOSHE KRAUSZ AND TIMOTHY OLENO
University of California, Berkeley

Two levels in an organization (supervisory level and subordinate level) independently participated in the development of behaviorally anchored rating scales. Results indicated that though the two organizational levels identified similar behavioral dimensions, the behavioral incidents defining the dimensions were valued differently. Alternative explanations for the obtained differences are offered and a discussion of behavioral expectation scale methodology is presented.

Several researchers concerned with the assessment of performance (Campbell, Dunnette, Arvey, & Hellervik, 1973; Fogli, Hulin, & Blood, 1971; Zedeck & Baker, 1972) have used behaviorally anchored rating scales that are developed through a procedure reported by Smith and Kendall (1963). The procedure is an iterative technique that results in scaled expectations of independent performance behaviors.

An aspect of the behavioral expectation scale (BES) procedure which has been somewhat ignored is that of interrater agreement for obtained scale scores. Though developers of behavioral expectation scales have argued that high-scale reliabilities are evidence of the value of behavioral expectation scales, Fogli et al. (1971) and Zedeck and Baker (1972) have indicated that the reliability estimation procedure is inappropriate and artifactual. A more appropriate reliability index is interrater agreement; that is, the agreement between two or more evaluators for a distribution of scale scores should be assessed when the scales have been applied in an evaluation situation. Two studies that have examined interrater agreement of specific behavioral expectation scales (Campbell et al., 1973; Zedeck & Baker, 1972) found low to moderate correlations (*r*s ranged from .24 to .55).

Zedeck and Baker (1972) also assessed the discriminant validity of behavioral expectation scales when used for evaluations; that is, the empirical independence of the scales was examined after the development procedure yielded conceptually independent scales. In general, results indicated poor discriminant validity.

Zedeck and Baker presented several alternative explanations for their obtained results. First, and most importantly, was the fact that the two supervisory evaluator groups, head nurses and supervisors, were at different organizational levels in the organizational hierarchy, each with different functions, responsibilities, and perspectives. It is possible that different supervisory groups differ in their opportunity to observe and evaluate subordinates.

A second possibility is that the different supervisory groups valued or perceived *different* behaviors as important in terms of meeting job requirements. If hierarchical position has some effect on the kinds of behaviors *expected*, furthermore, one could anticipate differences in behavioral expectations between not only two supervisory groups representing different organizational levels but also between a supervisory group and a subordinate group.

A third alternative is that two different supervisory groups can yield correlated ratings only when the behaviors used as anchors are valued similarly. That is, the two supervisory groups in each study may have valued the same behaviors differently. There is no evidence that the value placed on some behavior by one group of employees is the same as the value placed on the same behavior by a different group.

In short, there is reason to believe that a subordinate group may differ from a supervisory group regarding the behavior believed to be necessary for adequate performance. The possibility exists that there are value differences in expected and specific job behaviors, as indicated in the development of behavioral expectation scales, between two levels in an organization (supervisory level and a subordinate level). Although it was possible that different behavioral dimensions or qualities would be identified by the two groups, the focus in the present study was on the expectation that similar specific behaviors would be valued differently by the two levels. More exactly, it was anticipated that the subordinate group would rate a given set of behavioral examples as indicative of better performance than would a supervisory group.

---

[1] Requests for reprints should be sent to Sheldon Zedeck, Department of Psychology, University of California, Berkeley, Berkeley, California 94720.

METHOD

Subjects

The research site was a large, private university medical center in Northern California. The specific nursing specialties examined were operating room and recovery room nursing care. The supervisory group was composed of eight head nurses and the subordinate group was composed of nine registered nurses (RN). The same nurses participated throughout the study. In the last stage of the procedure, however, only six of the nurses in each of the groups participated.

Procedure

A modified version of the procedure reported by Smith and Kendall (1963) was used. The first stage required the determination of general behavioral dimensions on which nurses are to be evaluated. Several conferences were held independently with the head nurses (*n*s ranged from 6 to 8) and the RNs (*n*s ranged from 8 to 9) to achieve the objective. Once general dimensions were agreed upon, specific examples of behavior in each dimension were provided and edited into the form of expectations of specific behavior. Each nurse submitted three examples for each dimension; one example reflected good performance; one reflected satisfactory performance; and one reflected poor performance.

The second stage required that members of both groups indicate, independently, the dimension that was illustrated by each example obtained in the first stage. Examples that were not reallocated to the dimensions for which they were provided were eliminated. The criterion for retention of an example was 75% agreement. That is, 8 of the 12 nurses, with a minimum of 4 from each group, had to indicate that the example illustrated the same behavioral dimension. If less than 8 placed a specific behavior in the same dimension, that behavioral example was eliminated.

The third stage required that all remaining examples be assigned numerical values which reflected the desirability of the behavior illustrated. That is, each example was assessed on a scale ranging from 1 (poor) to 7 (good) by each of the nurses (RNs—$n = 6$; head nurses—$n = 6$). Means and standard deviations were computed for each example. Examples with standard deviations greater than 1.50 were eliminated.

RESULTS AND DISCUSSION

The first stage, the identification of behavioral dimensions, was conducted independently in the two groups. Several conferences with the head nurses resulted in the identification of [illegible] dimensions; the RNs also identified a set of 17 dimensions. There were 12 overlapping or similar dimensions. Five dimensions were identified only by the head nurses and 4 only by the RNs. The question arose as to whether the obtained differences were due to either substantial differences in job perspective or due to chance differences precipitated by inattention or oversight. Additional conferences were held to insure that the latter possibility did not account for the obtained differences. The role of the investigators who participated in both conferences with the head nurses and the RNs was merely informative and, thus, they suggested additional qualities to each group. The purpose of the conferences was to elicit a comprehensive description of the operating room and the recovery room nursing position and not to impose the values or assessment system of the investigators on the participants. These additional meetings resulted in common agreement that 22 dimensions ought to be considered in evaluating nursing personnel. The dimensions were as follows: adaptability; appearance; circulating ability; clinical appraisal; clinical judgment; communications—patient directed; communications—co-worker directed; cooperation and attitude; dependability, initiative; knowledge—aseptic technique; leadership ability; mechanical ability; organizational ability; operating room and recovery room orientation; perceptiveness; professional conduct; professional growth; quality; reaction to and function under stress; scrubbing ability; and teaching ability.[2]

The 22 dimensions were presented to members of both groups. After each nurse submitted three examples for each dimension, the result was 420 *different* examples of behavior. In the second stage, each nurse reallocated each of the 420 examples. Thus, each nurse evaluated examples submitted by her peers and by members of the other organizational levels. Elimination of examples which did not meet the 75% criterion resulted in an example pool of 249 examples and 21 dimensions. The quality dimension was eliminated since all examples provided for it were reallocated to other dimensions.

---

[2] Definitions of each dimension may be obtained from the first author.

The sign test (see Siegel, 1956, pp. 68–75) was used to analyze the data obtained in the third stage, that is, the assignment of numerical values to each of the 249 examples. The purpose of the sign test was to analyze the extent of the difference in the evaluation of the behavioral examples between the head nurses and RNs. Of the 249 examples evaluated, 177 examples were assigned higher mean values by the RNs (71%), the subordinate group, whereas 72 examples were assigned lower or equal values by the head nurses, the supervisory group. Note that this latter situation of 72 "less than" or "equal" evaluation is a slight departure from the usual sign test analysis. Ties or equals are usually omitted from the analysis yielding a reduction in the total *N*. In this study, ties were counted as "minuses" and worked against the hypothesized results; thus the conservative sign test was even more conservative employed in this fashion. Nevertheless, the frequency of higher valued examples by the RNs was significantly different ($p < .001$).

An analysis of mean values within each of the 21 dimensions indicated that the overall 71% (177/249) greater mean value assignment by the RNs was achieved for all but three dimensions: appearance, mechanical ability, and perceptiveness. For the quality of the appearance dimension, five examples were assigned higher mean values by the RNs, whereas seven examples were assigned higher mean values by the head nurses. For the mechanical ability dimension, seven examples were evaluated higher by the RNs, whereas eight examples were evaluated higher by the head nurses. For the perceptiveness dimension, three examples were evaluated higher by the RNs and three examples were evaluated higher by the head nurses. Examination of the individual examples in these three dimensions provided no reasonable hypotheses for the departure from the overwhelming result of greater mean value assignment by the RNs, the subordinate group. For the 18 dimensions within which the differences between the RN and the head nurse in mean values support the hypothesis, the percentages of higher value assignment by the RNs ranged from 64% to 92%.

CONCLUSIONS

The main aspect of performance evaluation investigated in this study was whether there are differences in the values assigned for a given behavior by two levels in an organization. Results for this study indicate that there are. Parenthetically, it might be noted that similar dimensions were identified by both the RNs and the head nurses, although more stringent controls should be applied in the future.

Head nurses, the supervisory level, overwhelmingly evaluated expected behavior of RNs, the subordinate level, to be less good than did the RNs themselves. This result may not be surprising if we accept the notion that supervisors expect more from their subordinates regardless of what the subordinates actually do. The result is surprising if we examine the nature of the position, nursing care in operating and recovery rooms. The impressions conveyed to the investigators by the head nurses and the RNs was that only those nurses who exemplify high-standard nursing proficiency are employed in the medical center studied. Exceptionally high standards or high proficiency, in addition, are expected in the operating and recovery room. The problem is that the different organizational groups had different high standards.

Are expected standards of performance determined by position in the organizational hierarchy? The two variables are closely related in this research sample. To answer this question properly, however, the present research methodology should be extended up and down the hierarchy. For example, groups participating could be nursing school students, beginning nurses in hospitals, experienced nurses, head nurses, supervisors, and, finally, administrators.

If differences with respect to expected standards are found, then the data obtained from the behavioral expectations procedure should be used for training purposes. When behaviors are identified for which there is disagreement in assigned values, conferences should be held for the purposes of discussing and resolving the differences.

One methodological problem in this study was that very few examples depicting satisfactory or average performance on a poor to good continuum were identified. This problem was recognized by other researchers using the behavioral expectations scale (Harari & Zedeck, 1972; Landy & Guion, 1970). Landy and Guion (1970) provided examples with values of poor and good performance for each dimension and then requested that participants assign values to the remaining examples. The anticipated result that more intermediate performance examples would be retained was not achieved.

This study specifically requested that participants indicate average or satisfactory performance. Nonetheless, few examples of average value were retained. A strategy for future investigations might be to provide specific anchors with values for poor performance, good performance, *and* average performance prior to assignment of values to the remaining examples. That is, specific behavioral examples which participants agree represent poor, good, and average performance should be identified in conferences. Then, with these as frames of references, values would be assigned to the remaining examples. It is anticipated that more intermediate examples would be obtained. If this did not occur, at least the one anchor of average value, developed in the conference, would be usable.

REFERENCES

CAMPBELL, J. P., DUNNETTE, M. D., ARVEY, R. D., & HELLERVIK, L. V. The development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology*, 1973, **57**, 15–22.

FOGLI, L., HULIN, C. L., & BLOOD, M. R. Development of first-level behavioral job criteria. *Journal of Applied Psychology*, 1971, **55**, 3–8.

HARARI, O., & ZEDECK, S. Evaluation of university faculty teaching. Paper presented at the meeting of the Western Psychological Association, Portland, Oregon, April 1972.

LANDY, F. J., & GUION, R. M. Development of scales for the measurement of work motivation. *Organizational Behavior and Human Performance*, 1970, **5**, 93–103.

SMITH, P. C., & KENDALL, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, 1963, **47**, 149–155.

SIEGEL, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.

ZEDECK, S., & BAKER, H. T. Nursing performance as measured by behavioral expectation scales: A multitrait-multirater analysis. *Organizational Behavior and Human Performance*, 1972, **7**, 457–466.

(Received February 8, 1973)

ERRATUM

In the article by Joseph Weitz and Seymour Adler, "The Optimal Use of Simulation," which appeared in the October 1973 issue (volume 58), a sentence on page 223 was incorrectly set. It should read: If, on the other hand, the very first trial on the real task is of great consequence (such as in emergency ditching procedures), then simulation training may be worthwhile.

JOURNAL OF APPLIED PSYCHOLOGY

Copyright © 1974 by the American Psychological Association, Inc.

June 1974 Vol. 59, No. 3

ARTICLES

Racial Differences in the Impact of Mathematics Training at a Manpower Development Program *Michael E. Gordon, Richard D. Arvey, William C. Daffron, Jr., and Dennis L. Umberger* 253

Identifying Responsive Inner-City Policemen *Rita Mae Kelly and Martin G. Farber* 259

Equal Unpleasantness Matching of Malodorous Substances in the Community *Thomas Lindvall and Leif T. Svensson* 264

Evaluation of a Deceleration Warning Light for Reducing Rear-End Automobile Collisions *John Voevodsky* 270

Effects of Different Levels of Integration on Advertising Preference and Intention to Purchase *George J. Szybillo and Jacob Jacoby* 274

A Dynamic Model of Relationships among Advertising, Consumer Awareness, Attitudes, and Behavior *David A. Aaker and George S. Day* 281

Components of Perceived Risk in Product Purchase: A Cross-Validation *Leon B. Kaplan, George J. Szybillo, and Jacob Jacoby* 287

Effect of Anonymity on Return Rate and Response Bias in a Mail Survey *Carol Fuller* 292

Contribution of Personalization to Mail Questionnaire Response as an Element of a Previously Tested Method *Don A. Dillman and James H. Frey* 297

Scaling Peer Ratings: An Examination of the Differential Validities of Positive and Negative Nominations *Gary G. Kaufman and James C. Johnson* 302

An Empirical Comparison of the Relative Effects of Rater Response Biases on Three Rating Scale Formats *Robert F. Burnaska and Thomas D. Hollmann* 307

Magnitude Estimations of Expressions of Frequency and Amount *Bernard M. Bass, Wayne F. Cascio, and Edward J. O'Connor* 313

Measuring Organizational Communication *Karlene H. Roberts and Charles A. O'Reilly III* 321

Factors Influencing Disciplinary Judgments *Benson Rosen and Thomas H. Jerdee* 327

Effects of Incentive, Feedback, and Manner of Presenting the Feedback on Leader Behavior *Richard P. Butler and Cabot L. Jaffee* 332

Participative Management in the College Classroom *Ralph H. Kilmann* 337

Intrinsic, Extrinsic, and Risk Propensity Factors in the Work Attitudes of Teachers, Educational Administrators, and Business Managers *Cecil Miskel* 339

Cluster Analyses of Interest Inventory Scales as Tests of Holland's Occupational Classification *Dean H. Nafziger and Samuel T. Helms* 344

Checking Groups of Letters *Ian E. Gordon and Moira Amos* 354

Effectiveness of Highlighting for Retention of Text Material *Robert L. Fowler and Anne S. Barker* 358

A Factor Analysis of Air Passenger Reactions to Skyjacking and Airport Security Measures as Related to Personal Characteristics and Alternatives to Flying *Michael R. Cooper, Charles E. Boltwood, and Robert J. Wherry, Sr.* 365

SHORT NOTES

Replication of White-Collar–Blue-Collar Differences in Sources of Satisfaction and Dissatisfaction *Thomas C. Harris and Edwin A. Locke* 369

Correlates of Job Satisfaction in Naval Environments *Blair W. McDonald and E. K. Eric Gunderson* 371

(*Continued on next page*)

886 (3)
16 | 9 | 74

(Continued from previous page)

Relationships Between Performance and Satisfaction Under Contingent and Noncontingent Reward Systems
Gerald A. Kesselman, Michael T. Wood, and Eileen L. Hagen 374

Relationship of Early Work Challenge to Job Performance, Professional Contributions, and Competence of Engineers *H. G. Kaufman* 377

Some Moderator Influences on Relationships Between Consideration, Initiating Structure, and Organizational Criteria *Irwin J. Badin* 380

A Note on the Predictive Validity of Initiating Structure and Consideration in ROTC Training *M. M. Petty and Norman M. Pryor* 383

Relationship Between Supervisory Cognitive Style and Social Orientation
L. W. Gruenfeld and P. Weissenberg 386

Dogmatism and the Congruence Between Self-Reported Job Preference and Performance Among School Supervisors *James P. Esposito and Herbert C. Richards* 389

Unity of Command and Job Attitudes of Managers in a Bureaucratic Organization
Martin J. Gannon and Frank T. Paine 392

Repetition of Highly Similar Messages and Attitude Change
J. Lee McCullough and Thomas M. Ostrom 395

Development of Scales to Measure Attitudes Toward Marijuana and Marijuana Users
Victoria Colaiuta and George Breed 398

Landscape Office Attitudes: Reflections of Perceived Degree of Control Over Transactions with the Environment
Michael W. McCarrey, Lyle Peterson, Shirley Edwards, and Paul Von Kulmiz 401

Development of a Biographical Inventory Blank to Predict Faking on Personality Tests
Jeffrey Cohen and Joel Lefkowitz 404

The Influence of Technical Knowledge on Position Error in Ranking
Edwin E. Wagner and Thomas O. Hoover 406

Brightness Contrast Effects on Recall of Projected Highway Sign-Type Stimulus Material *Robert W. Sweezey* 408

LIST OF MANUSCRIPTS ACCEPTED 263, 336

SPECIAL ANNOUNCEMENT 400

INSTRUCTIONS TO AUTHORS

Articles must possess: (*a*) conciseness; (*b*) unambiguous and simple vocabulary; (*c*) conformity to accepted technical style in tables, terminology, and references; and (*d*) conclusions that are clearly related to the evidence presented. Authors must progress from a statement of problem or purpose, through analysis of evidence, to conclusions and implications.

Criteria for evaluation include: (*a*) significance of contribution, (*b*) technical adequacy, (*c*) appropriateness for the *Journal of Applied Psychology,* and (*d*) clarity of presentation.

See the *Publication Manual of the American Psychological Association* (1967 Revision), $1.50, for the format of manuscripts. Articles not prepared in this manner cannot be reviewed. Note sections on typing (p. 48, *all copy must be double-spaced*), tables, figures, and references. The first author's last name must be typed on the upper left-hand corner of each page, except for blind reviews. Do not abbreviate, for example, journal titles in reference lists, the word "subject," names of groups, etc. Each manuscript copy must include an abstract of 100–120 words typed on a separate sheet of paper.

Optional blind review must be specifically requested when a manuscript is first submitted. The author's name and affiliation must appear only on a separate title page, included with each copy of the manuscript. Footnotes containing the identity of the author or his affiliation must be on separate pages.

Check recent issues of the *Journal of Applied Psychology* for approximate length of regular articles. (One printed page equals three double-spaced manuscript pages.) A few longer articles of special significance are printed as monographs. Short Notes feature brief reports on studies such as those involving some methodological contribution or important replication.

The original drawing of a figure or an 8 × 10 inch glossy print must be submitted with the original copy of the manuscript. Duplicate copies may be photographic or pencil drawn. Figures *must* be hand-lettered.

Supplementary materials formerly deposited with the National Auxiliary Publications Service need no longer be submitted with the manuscript. Retain supporting and raw data for at least 5 years after publication of the article.

Journal of Applied Psychology
1974, Vol. 59, No. 3, 253–258

RACIAL DIFFERENCES IN THE IMPACT OF MATHEMATICS TRAINING AT A MANPOWER DEVELOPMENT PROGRAM [1]

MICHAEL E. GORDON [2] AND RICHARD D. ARVEY

University of Tennessee

WILLIAM C. DAFFRON, JR., AND DENNIS L. UMBERGER

Tennessee Eastman Company, Kingsport, Tennessee

The impact of mathematics instruction on the measured numerical competence of a class of 80 black and 83 white trainees was investigated at a successful manpower development project. The California Achievement Test—junior high school level—was administered at successive six-week intervals over a six-month period of time. The results revealed a significant Race × Training interaction which indicated that the whites profited more from the instruction than did the blacks. These findings imply that racial differences in measured mathematical competence are not reduced by training but may in fact be increased.

The provision of equal educational opportunities and special training programs for disadvantaged segments of the population is frequently offered as a means of reducing discrepancies in measured cognitive competence and in so doing have an impact on the limitations in vocational achievement that these discrepancies impose. The large number and variety of publicly and privately sponsored programs testify to the enthusiasm with which remedial or compensatory instruction has been embraced as a means of reducing these disparities. Special educational programs exist for children of all ages, from the widely publicized preschool Head Start program, through elementary school programs such as Project Concern (Mahan, 1968), to high school and college level programs such as Project ABC (Wessman, 1969). Industrial training for the hard-core unemployed and the disadvantaged also has stressed instruction in basic skills designed to upgrade the trainee's ability in English and mathematics.

With regard to racial differences, most of the evidence on the effects of special education on measured differences in performance on cognitive tests is cross-sectional research and, unfortunately, does not provide a direct test of the impact of training on the relative numerical and verbal skills of black students and white students (for a recent review of the impact of special educational programs, see St. John, 1970). The few longitudinal validation attempts have been directed at programs for very young children (kindergarten through second grade) and, typically, have indicated that initial racial differences in cognitive competence are reduced because of the special educational opportunities provided (e.g., Katzenmeyer, 1963; Stallings, 1959).

Precious little longitudinal research has been performed to determine the effect of industrial training programs on racial differences in test performance between mature blacks and whites. A review of recent training literature revealed only one relevant study (Miller, 1971) that attempted to monitor the effects of a basic education program on adult test performance by eight Caucasians and eight Chippewa Indians.

There are several important reasons which necessitate the study of the impact of industrial training on the intellectual competence of blacks vis-à-vis whites. First, it is important to examine whether either group (blacks or whites) profits from a given learning experience in order to ensure that the training program is having a desired effect. Second, if differential training effects are found between blacks and whites, quite possibly different training techniques and procedures might be implemented. Third, it is useful to examine any differences that may exist between racial groups rather than to ignore or attempt to mask them. To ignore differences that may actually exist could result in sponsorship of training which exacerbates the problem of racial differences.

---

[1] Support for this study was secured by a grant from the U.S. Department of Labor and the Atomic Energy Commission under interagency agreement (DOL 82-45-69-09, AEC 40-57-68).

This report is based in part on the theses prepared by the third and fourth authors, under the direction of the first author, in partial fulfillment of the requirements for the master's of science degree in industrial management at the University of Tennessee.

[2] Requests for reprints should be sent to Michael E. Gordon, Department of Industrial and Personnel Management, University of Tennessee, Knoxville, Tennessee 37916.

The present research was designed to compare differences in test performance between blacks and whites before, during, and after exposure to a federally sponsored manpower development project for the unemployed and underemployed. Specifically, the trainability of black students and white students in the area of mathematics was investigated in a longitudinal study involving the use of standardized tests at six-week intervals over a period of six months.

METHOD

Research Site

The study was performed at a manpower development project located in eastern Tennessee that provided instruction in six industrial skills: machining, welding, mechanical operations, physical testing, electronics, and drafting. In addition to the skill training, related instruction was provided in communications and mathematics. The primary objective of the mathematics instruction was to provide the trainee with a ninth-grade high school level of competence. To this end trainees attended classes three times a week on the average to receive instruction in computations involving fractions and decimals, recognition of formulas and symbols, algebra, and trigonometry. The related instruction was based on the mathematics requirements of the particular industrial skill that each trainee was learning. Private tutorial sessions were also provided to help those trainees who were falling behind in the classroom sessions. Although no formal records were immediately available to substantiate what occurred in these private sessions, the supervisor of related instruction indicated that blacks were more frequently involved than were whites in these special tutorials.

The effectiveness of this particular training program has been well documented in terms of both economic criteria, for example, cost-benefit analysis (Kirby & Castagna, 1969), and behavioral criteria, for example, the effects on the personal lives of the trainees (Gordon & Scott, 1972). The project must be adjudged successful by comparison with other manpower development programs in terms of the high percentage of trainees (79% for the entire class of trainees involved in this study) who complete the training and are graduated.

Subjects

The subjects comprised the class which began its training in April 1970. The trainees were recruited from a four-state Appalachian region by a variety of agencies such as the Chattanooga Concentrated Employment Program. A total of 180 subjects representing all six training areas were tested on their entry into the program. Of this number, 80 were black and 83 were white. The black trainees and the white trainees were quite similar on the average with respect to age, education, and employment status on entry, but differed with respect to the percentage of trainees that completed the program (84% of the white trainees versus 75% of the black trainees completed the program).

Four American Indians and three Spanish-speaking trainees were also tested, but their data were ignored in subsequent analyses. Ten additional subjects whose race was not determined were also discarded.

Criterion

The two arithmetic subtests which comprise the junior high school level of the California Achievement Test (CAT; California Test Bureau, 1967) were used as measures of competence in mathematics. The Arithmetic Reasoning subtest consisted of 55 questions and was divided into three sections: meanings, worded problems, and symbols, rules, and equations. The Arithmetic Fundamentals Test consisted of 80 items and was subdivided into four sections: addition, subtraction, multiplication, and division. The criterion employed was the total score based on both subtests.

The four alternate forms of the CAT (W, X, Y, and Z) were used in this research. Unfortunately the published norms, reliabilities, and validities applied only to Form W.[3] Although they contain parallel content, the equivalency of the alternate forms has not been verified empirically.

Procedure

Premeasures. In order to test equivalency, each subject took two different forms of the CAT as part of the orientation program for new trainees. The 153 subjects who were present for the first day of orientation were randomly subdivided so that approximately equal numbers took one of the six possible pairs of the four forms. Equivalency was checked by comparing the means and variances of the four forms and by comparing the magnitude of the six intercorrelations among all forms of the test. Separate normative data on blacks and on whites

---

[3] All subtests of Form W have reported reliabilities about .80.

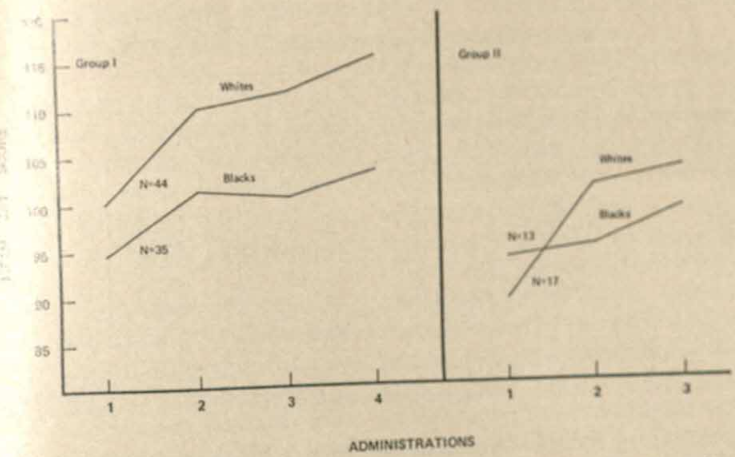


FIGURE 1. Mean California Achievement Test (CAT) scores for blacks and for whites for each administration in Groups 1 and 2.

were not prepared due to the limited sample size of each group.[4]

Bartlett's test was used to determine whether the variances were homogeneous. The observed chi-square of 1.64 ($df = 3$) was not significant ($p > .10$) which indicated that it was tenable to accept the null hypothesis of equivalent variances among the alternate forms.

A one-way analysis of variance was performed to test the significance of the difference among the means of the alternate forms. It appears as though the means of the forms were equivalent except for random variation, since the obtained F ratio of 1.59 ($df = 3/149$) was not significant ($p > .10$).

Finally, the homogeneity of the six intercorrelations among the four forms was determined using the chi-square procedure originated by Fisher (see Edwards, 1960). The intercorrelations were in the range between .93 and .96 except for the correlation between Forms X and Z which was .86. The observed chi-square of 5.45 ($df = 5$) was not significant ($p > .10$), a fact which suggested that the four forms could be considered equivalent since their intercorrelations were all estimates of the same population value.

Postmeasures. After the equivalency of the four CAT forms was demonstrated, three additional administrations were scheduled at six-week intervals over the training program in order to measure changes in arithmetic competence. The tests were administered in regularly scheduled math classes by the related instruction teacher. The subjects were given each of the four forms of the test at least once, although the order of administration was randomized for each trainee. Owing to limited resources, no effort was made to follow up on a trainee who was absent on the day a test was administered. Only those subjects who had complete longitudinal data covering a minimum of the first 12 weeks of training, that is, all four CAT administrations, were used in the study. Using these guidelines for inclusion, data were available for 35 black subjects and 44 white subjects on all five administrations of the test (Group 1). Another group of 13 black subjects and 17 white subjects had complete data on the first four tests (Group 2).[5] Overall, 60% of the black trainee population and 73% of the white trainee population were the basis for subsequent analyses.

RESULTS

The data from Groups 1 and 2 were evaluated separately, although the analytical procedures employed were identical in both cases. Figure 1 contains plots of the means on each of the administrations for each racial group. Note that Administration 1 in both graphs represents the average of the two tests given on successive days during orientation as premeasures. Examination of the standard deviation within racial groups for each trial did not reveal any important differences or trends either between or within groups with respect to variability of the criterion distributions.

The data summarized in Figure 1 were subjected to a two-factor analysis of variance with one repeated measure (Winer, 1962). The assumption of homogeneity of variance–

---

[4] Tables containing means, variances, and intercorrelations for each of the four California Achievement Tests (CAT) forms are available on request from the first author.

[5] This group was composed primarily of welding trainees who received just three months of math instruction.

TABLE 1

SUMMARY OF ANALYSIS OF VARIANCE FROM GROUPS 1 AND 2

Source of variation	Group 1			Group 2		
	MS	*df*	*F*	*MS*	*df*	*F*
Race (A)	6,943.85	1	5.11[a]	101.95	1	.09
Subjects within A	1,358.34	77		1,125.73	28	
Administrations (B)	2,180.21	3	44.48**	839.08	2	31.26***
A × B	156.33	3	3.19*	248.29	2	9.25***
B × Subjects within A	49.02	231		26.84	56	

[a] Significant at the $\alpha = .025$ level.
* Significant at the $\alpha = .10$ level ($df = 1/77$) based on the conservative test suggested by Greenhouse and Geisser (1959).
** Significant at the $\alpha = .01$ level ($df = 1/77$) based on the conservative test suggested by Greenhouse and Geisser (1959).
*** Significant at the $\alpha = .01$ level ($df = 1/28$) based on the conservative test suggested by Greenhouse and Geisser (1959).

covariance matrices on which this analysis of variance procedure is based was tested, and the observed chi-squares were significant in both groups ($p > .01$). Therefore, the more conservative test suggested by Greenhouse and Geisser (1959; see also Winer, 1962, p. 305) was employed to determine the significance of the F ratios obtained in the analysis of variance procedure involving the repeated measure (the administrations factor). These results are reported in Table 1.

Examination of the results revealed that the administrations and the Race × Administrations effects were statistically significant in both samples. The significant interaction terms prompted tests of simple main effects of race for the first and last administration in both groups. In Group 1 the differences between the means of the whites and the blacks were not significant ($p > .10$) on the premeasure ($F = 1.66$, $df = 1/77$) but they were significant ($p < .01$) on the fourth administration ($F = 7.54$, $df = 1/77$). In Group 2 the simple main effects were not significant ($p > .10$) for either the premeasure ($F = .36$, $df = 1/28$) or the third administration ($F = .36$, $df = 1/28$).

TABLE 2

INTERCORRELATIONS AMONG ADMINISTRATIONS IN BOTH GROUPS FOR EACH RACE

Administration	Group 1			Group 2	
	2	3	4	2	3
			Blacks		
1	.92	.89	.80	.97	.94
2		.92	.76		.93
3			.80		
			Whites		
1	.92	.85	.79	.93	.92
2		.94	.91		.97
3			.93		

Note. All of the reported correlations are significant at $\alpha = .01$.

Finally, the stability of the trainees' relative competence in math was examined by computing the intercorrelations among the results on each administration for each race separately. These results are reported in Table 2. All the observed correlations were significant ($p < .01$), and these results strongly suggest that the trainees maintained their relative standing within their racial groups throughout the training period.

DISCUSSION

The results from Groups 1 and 2 indicate a differential training effect for blacks and for whites. The significant main effects for the administrations suggested that both races profited from the training. However, the whites clearly derived greater benefit from the training experience than did the blacks, a fact revealed by the significant Race × Administrations interaction terms observed in both Groups 1 and 2. This differential training effect was especially prominent for Group 1 in which the instruction produced significant racial differences in terms of measured mathematical competence between groups of subjects that appeared alike on the premeasure.

Before the differential training effects observed are attributed to different racial levels of ability to learn, a number of potentially confounding variables must be examined. Three such variables that suggest viable alternative explanations for the results are discussed below.

First, the role of socioeconomic status as a determinant of learning ability has received a great deal of attention in educational literature (Jensen, 1968). It has been noted frequently that low-socioeconomic-status children perform more poorly on cognitive tests than do high-socioeconomic-status children and that there is a significant correlation between educational achievement and socioeconomic status. In order to rule out the possibility that differences in socioeconomic status might have been related to the differential training effect observed, additional background data on the subjects were scrutinized. Examination of the limited socioeconomic status information available on the subjects in both Groups 1 and 2 revealed that 94% of the blacks and 61% of the whites were classified as "disadvantaged."[6] Learning curves were prepared for Groups 1 and 2 on the basis of this measure of socioeconomic status.

The 42 disadvantaged subjects in Group 1 were composed of 20 whites and 22 blacks, while the 17 advantaged subjects were composed of 15 whites and 2 blacks. These 59 subjects for whom socioeconomic status data were available represented approximately 75% of the original 79 subjects in Group 1. The same 2×4 analysis of variance procedure used previously was applied to Group 1 data subdivided on the basis of socioeconomic status. The results indicated that the main effect for socioeconomic status was not significant ($F = 1.85$, $df = 1/57$, $\alpha = .10$). Consequently, socioeconomic status probably may be disregarded as an important variable in this research.[7]

Second, a closer look at the nature of the learning environment also is necessary in order to afford a clearer interpretation of the results previously presented. Obviously, if the black trainees did not receive treatment and instruction of the same caliber as that received by the white trainees, the differential effects could be explained easily in terms of environmental factors in the training situation. However, it is very difficult to collect and present objective data on the parameters of a learning situation that might reveal the existence of either persistent or intermittent racial discrimination or favoritism. Consequently, the only evidence regarding the racial climate that can be offered is anecdotal. The first author acted as a consultant on personnel problems and research for two years at the project described in the present study. During this time, he had ample opportunity to observe the staff of trainers which appeared both dedicated and egalitarian. Although there were occasional reports of racial problems at the project, close scrutiny of the data available on training personnel and trainees failed to substantiate any of these isolated reports. Certainly there was no systematic or pervasive discrimination against blacks that might be offered as a viable explanation for the observed differences in training effects. Furthermore, the high stability of measured competence within the black sample over the duration of the training militates against an explanation of the lower CAT performance on the basis of frequent incidents of discrimination against randomly selected trainees (the median coefficient of stability for the six-month time interval was .80 for the blacks in Group 1).

A third relevant variable which deserves attention is the motivation to perform well on tests and/or the will to achieve in the training program. It is possible that the blacks perceived little reason for exerting themselves on the tests or in the classroom because they saw few, if any, rewards contingent on high performance. Certainly, the historical record of discrimination in hiring practices in the United States tends to develop and support lower expectation among the black trainees. Unfortunately, data on expectancies and perceived rewards were not available on these subjects.

---

[6] The term "disadvantaged," according to the U.S. Department of Labor guidelines, is defined as follows: "A disadvantaged individual, for manpower purposes, is a poor person who does not have suitable employment and who is either (1) a school dropout; (2) a member of a minority; (3) under 22 years of age; (4) 45 years of age or over; or (5) handicapped."

[7] A similar analysis on Group 2 was not attempted given the limited number of advantaged subjects identified.

Although these results suggest that industrial training programs may be actually increasing the disparity in measured competence between black trainees and white trainees, one must not overlook the benefits which accrued to members of both racial groups. A ninth-grade level of competence in math was attained by approximately 90% of the white trainees and approximately 80% of the black trainees. Furthermore, there was nothing in these results which suggested that the mathematical learning in black subjects had reached a final asymptote. More concentrated or protracted training of blacks relative to whites might have caused the former trainees to reach the same level of competence as the latter. However, it is of substantial importance that trainers recognize the possibility that the instruction they provide may have to be modified in order to avoid the differential training effects noted in the present study.

These considerations suggest the need for further research to replicate and extend these findings. A rigid interpretation of the results in terms of learning ability certainly is not warranted at this time. Motivational elements which might account for these data should be explored as should the relationships of these test differences to future job performance criteria.

Finally, a related methodological problem has been raised by these results, namely, the possibility of the CAT unfairly discriminating against black applicants. Although the guidelines for considering a test unfair to blacks are not entirely clear, one standard index typically examined in the determination of the fairness of a test is the mean difference between racial groups (Bartlett & O'Leary, 1969). It is interesting to note that in terms of the results for Group 1 the fairness of the CAT would not be questioned for the untrained subjects but would create serious misgivings at the conclusion of training. Clearly, the absence of a job performance criterion in the present study prohibits a discussion of the validity of the CAT for the two racial groups. However, given a selection ratio of less than .75, any company using the CAT to select graduates of the program for job entry level employment would surely hire substantially more whites than blacks.

REFERENCES

BARTLETT, C. J., & O'LEARY, B. S. A differential prediction model to moderate the effects of heterogenous groups in personnel selection classification. *Personnel Psychology*, 1969, 22, 1–17.

EDWARDS, A. L. *Experimental design in psychological research*. New York: Holt, Rinehart & Winston, 1960.

GORDON, M. E., & SCOTT, R. D. Evaluation of a manpower development project in terms of its effects on the personal lives of its graduates. *Journal of Vocational Behavior*, 1972, 2, 467–478.

GREENHOUSE, S. W., & GEISSER, S. On methods in the analysis of profile data. *Psychometrika*, 1959, **24**, 95–112.

JENCKS, C. Social stratification and higher education. *Harvard Educational Review*, 1968, **38**, 277–316.

KATZENMEYER, W. G. Social interaction and differences in intelligence test performance of Negro and white elementary school pupils. *Dissertation Abstracts*, 1963, **24**, 1904–1905.

KIRBY, F. C., & CASTAGNA, P. A. *Benefit-cost analysis of TAT Phase I worker training*. (Special Report, TAT Project) Oakridge, Tenn.: Oakridge Associated Universities, July 1969.

MAHAN, T. W. *Project Concern—1966–1968: A report on the effectiveness of suburban school placement for inner-city youth*. Hartford, Conn.: Board of Education, August 1968.

MILLER, B. P. IQ tests and minority groups. *Training and Development Journal*, 1971, 25, 26–27.

ST. JOHN, N. H. Desegregation and minority group performance. *Review of Educational Research*, 1970, **40**, 111–133.

STALLINGS, F. H. A study of the immediate effects of integration on scholastic achievement in the Louisville public schools. *Journal of Negro Education*, 1959, **28**, 439–444.

CALIFORNIA TEST BUREAU. *Technical report on the California Achievement Tests, 1963 Norms*. Monterey: Author, 1967.

WESSMAN, A. E. *Evaluation of Project ABC (A Better Chance): An evaluation at Dartmouth College Independent Schools Scholarship Program for Disadvantaged High School Students*. (U.S. Department of Health, Education and Welfare; Office of Education) Washington, D.C.: U.S. Government Printing Office, April 1969.

WINER, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

(Received March 5, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 259–263

IDENTIFYING RESPONSIVE INNER-CITY POLICEMEN[1]

RITA MAE KELLY[2] AND MARTIN G. FARBER

American Institutes for Research, Kensington Office, Maryland

Police responsiveness was assessed via a technique identifying positive and negative attitudes of policemen towards inner-city residents. Analysis of some demographic characteristics of the most positive (responsive) and most negative (nonresponsive) groups revealed that race was the most important distinguishing variable between them. Evaluations of the behavior of these officers by their immediate superiors supported the assumption that police in the positive group are rated higher than those in the negative group.

A critical social problem in inner-city ghetto areas is the relationship between police and residents. It has been considered essential that police officers should be responsive to community needs (Clark, 1970). However, there are differences of opinion about both the definition of a good, responsive policeman, and the best method of recruiting and training him to maximize responsiveness. Police officials tend to believe that a good policeman is a good policeman, regardless of race, color, creed, or cultural background (Bouma, 1969; Radelet, 1966; Wilson, 1968). Conversely, community leaders tend to argue that responsiveness is best found among policemen who are similar to the community in terms of race, place of residence, and cultural upbringing (Edwards, 1968; Mast, 1970).

The recruitment of minority group members as policemen and the effect of that recruitment on police forces and communities required exploration. Earlier studies (Alex, 1969; Hunt & Cohen, 1971) reached differing conclusions about whether minority group policemen are as service oriented as white policemen. Kelly and West (1972), comparing 193 black and 316 white Washington, D.C., policemen, found that a higher proportion of blacks than whites were service oriented. They also found, contrary to popular assumptions, that the transition of the Washington, D.C., police force from one with almost no blacks to one with 38% black officers did not change the general socioeconomic class background of the total force.

The research reported in the present study sought to develop an operational definition of "responsiveness." It addressed the policy questions of: (*a*) If responsive policemen are recruited, what are the chances of getting blacks as compared to whites? (*b*) If blacks are specifically recruited, what are the chances of getting responsive policemen? This study presents analytic procedures for developing an empirical definition of responsive policemen and explores the relationship of this empirical definition to the evaluation of policemen by their superiors. Finally, some initial results of analyses of the connection between demographic variables and the most and least responsive police groups are presented.

METHOD

Subjects

A 10% random sample (stratified by rank and extent of contact with the community) was chosen from the District of Columbia Metropolitan Police Department. Eighty-five percent (509 policemen) completed usable questionnaires. The response rate for higher ranking officers was 85%, for lower ranking officers it was 84%. Variation in response rate by extent of contact was also small, ranging from 84% to 90%, with those having least direct street contact with residents having the highest response rate. Of the 509 respondents, 38% were black, the exact percentage of blacks on the police force at that time.

---

[1] The data used in this article was collected and analyzed under contract BOO-5133, between the Office of Economic Opportunity and the American Institutes for Research. The opinions expressed herein are those of the authors and should not be construed as representing the opinions or policy of any agency of the United States government or of the American Institutes for Research. Gratitude is expressed to both institutions for permission to use the data.

[2] Requests for reprints should be sent to R. Kelly, 10 Wilson Terrace, Elizabeth, New Jersey 07208.

Analysis of four Guttman scales (see Design) revealed that those policemen who scored negatively on the Fear of Citizens Scale (the Police Defensive Posture Scale) also tended to score negatively on the other three Guttman scales. Therefore, a card sort was done on the 509 subjects to determine how many scored positively on all scales and vice versa. This study is based on the sample of the 30 policemen who scored negatively and the 32 who scored positively on all four scales. These 12% of the original 509 subjects could be classified as "extremes." It was felt that the scales might have tapped a basic attitudinal orientation toward responsiveness or nonresponsiveness. Therefore, the 32 most positive group was designated "responsive," and the 30 most negative group was called the "nonresponsive" group.

Design

A portion of the questionnaire contained attitude scales designed to elicit responses regarding police attitudes toward and relationship with residents of their work areas (Kelly, Dockett, Farber, West, & Gray, 1972). Five of Guttman's criteria of unidimensionality were applied (Stouffer, et al., 1950): (*a*) total sample and individual item reproducibility near .90, (*b*) an adequate range of marginal frequencies, (*c*) lack of substantial nonscale items, (*d*) at least four items per scale, and (*e*) sufficiently numerous and varied response categories. Because of time and expense, the last criterion was not strictly applied. The original categories (strongly agree, agree, neutral, disagree, strongly disagree) were collapsed into positive/negative dichotomies with neutrals being considered positive. Using these criteria, four scales were identified.

The first scale, Policemen's View of Citizens' Support, asked how often residents (*a*) take a hand when juveniles and children are causing trouble or are on the verge of breaking the law; (*b*) identify themselves when necessary and appear in court when requested; (*c*) help police identify criminals; and (*d*) report crimes they observe.

The second scale, General Faith in People, inquired about the extent of agreement with the statements: (*a*) No one is going to care much what happens to you, when you get right down to it. (*b*) Human nature is fundamentally cooperative. (*c*) If you don't watch yourself, people will take advantage of you. (*d*) Most people can be trusted.

The third scale, Police Fear of Citizens or Police Defensive Posture, was measured by intensity of agreement/disagreement on five items: (*a*) In this city, physical combat skills and an aggressive bearing are more useful to a patrolman on the beat than book learning and a courteous manner. (*b*) Most people in my district do not respect policemen. (*c*) Residents in this city view the police as an alien-occupying army. (*d*) A policeman must be on guard or residents will take advantage of him. (*e*) The likelihood of a policeman being abused by citizens in this city is very high.

The fourth scale, Police Stereotype of Poor People, was composed of these items: (*a*) If the truth were known about poor people, it is that they are lazy and don't really want to work. (*b*) People live in poverty areas only because they were unwilling to help themselves. (*c*) All people in poverty areas want is a handout without working for it. (*d*) One of the main causes of poverty is lack of moral strength and willpower.

RESULTS

Responsiveness and Supervisor's Ratings

The two extreme groups were compared to determine how the operational definition of responsiveness might be related to the policemen's actual behavior. Comparisons were based on ratings of policemen by their superiors in the areas of "bearing and behavior" (the extent to which the ratee meets standards of bearing, dress, courtesy, and emotional control, and enhances the image of the police department) and "human relations" (the degree to which the ratee promotes harmonious relationships with other members of the police department and with civilian associates). As expected (see Table 1), officers in the positive group rated well above their negative counterparts on both dimensions. Compared to only 30% of the negatives, 69% of the positives were rated above average in bearing and behavior. (Note that 40% of the negatives were under 25 years of age and were too new on the force to have had ratings placed in their files.) On the human relations dimension, 72% of the positives were rated above average, compared to only 33% of the negatives.

An overall rating for each policeman considered both these dimensions and such things as learning ability, knowledge and skill, and acceptance of responsibility. The two groups showed no statistically significant difference in this study, but more of the positives than the negatives were rated above average.

This correlation between ratings and responsiveness group indicates that the Guttman scales are assessing both a basic attitudinal orientation and a strong behavioral predisposition.

Responsiveness and Demographic Characteristics

Important social questions are: Will better educated policemen be more responsive in inner city areas? Will black policemen be more responsive? Although the particular data analyzed in this study do not answer these questions, they do permit an assessment of the question: What proportion of the policemen within each of the extreme groups have particular demographic characteristics? So that the reader can judge what potential connection the defined categories of responsive policemen might have to these direct policy questions, some brief analyses are discussed below. (See Table 2.)

1. *Race*. All 30 policemen in the most negative group are white (two had Spanish surnames). Of the 32 most positive respondents, 56% are white. Race thus seems quite significant in determining the extent to which policemen perceive the inner-city residen negatively, but does not seem as strongl associated with the composition of the positive group.

2. *Education*. It is often asserted that individuals with advanced education beyond hig school level make better policemen than thos with lower educational attainment. An analysis of level of education, however, indicate that the groups are almost identical in th

TABLE 1

RESPONSIVENESS AND SUPERVISORS' RATINGS

Supervisors rating	Least responsive ($n = 30$)	Most responsive ($n = 32$)
	Bearing and behavior[a] (%)	
Average or below	30	22
Above average	30	69
NA	40	9
	Human relations[b] (%)	
Average or below	27	19
Above average	33	72
NA	40	9
	Overall evaluation[c] (%)	
Average or below	27	25
Above average	33	65
NA	40	9

Note. NA = not applicable.
[a] p for Fisher's Exact one-tailed test = .067.
[b] p for Fisher's Exact one-tailed test = .081.
[c] p for Fisher's Exact one-tailed test = .20.

TABLE 2

RESPONSIVENESS AND SELECTED DEMOGRAPHIC CHARACTERISTICS

Demographic characteristics	Least responsive ($n = 30$)	Most responsive ($n = 32$)
	Race* (%)	
Black	0	44
White	100	56
	Education (%)	
No high school diploma	17	19
High school graduate	43	44
Some higher education	37	37
NA	3	0
	Rank (%)	
Officer	73	63
Higher ranks	27	37
	Age (%	
20–24	40	19
25–34	27	38
35 plus	33	43
	Place of residence* (%)	
Washington, D.C.	17	53
Maryland	60	22
Virginia	23	25

Note. NA = no answer.
* $p < .01$ for χ^2.

distribution of the respondents within the various educational levels.

3. *Rank.* No sharp differences exist in terms of the rank of the policemen either, although the proportion of respondents with higher ranks in the most positive group was 10 percentage points greater than in the most negative group.

4. *Age.* Although the difference in age between the two groups is not statistically significant, the largest percentage of the most negative are 24 or younger, while the largest proportion of the most positive sample is 35 or older. New recruits and younger policemen might be more negative toward residents because (*a*) they tend to be more idealistic in their expectations; (*b*) they are embarking on a still unknown career at a time when the police profession is coming under great public scrutiny; or (*c*) recruitment criteria might be lower now than previously. Whether policemen's attitudes become more positive with time or those with the more negative attitudes drop out cannot be determined with these data.

5. *Place of residence.* A sharp difference exists with regard to where the policemen in the two groups reside. This significant difference is expected due to the direct relationship between race and place of residence within the Washington, D.C. police force (see Kelly & West, 1972). Only 17% of the most negative officers live in the city, whereas 53% of the most positive ones reside in the city which they serve. The effect of race on place of residence can be seen by the fact that of the 17 most positive respondents living in Washington, D.C., 13 (77%) are black. Conversely, of the 15 positive respondents living outside the city, 14 (93%) are white.

CONCLUSION

The data indicate that an empirical approach to identifying responsive/nonresponsive policemen is feasible. The fact that 12% of the original 509 policemen surveyed fell into one of the extreme groups indicates that the dimensions assessed by the four Guttman scales are real. Further, the results from demographic analyses reveal that the identification of the most positive/negative policemen by the Guttman scaling technique meets with popular conceptions of what a sensitive measure of this dimension would find.

Further research is needed to determine the precise policy implications of the above findings for police selection. The data in this study provide only a crude estimate of the answer to the policy question popularly phrased: If responsive policemen are recruited, what are the chances of getting blacks as compared to whites? Of the 32 most responsive policemen, 44% were black. The District of Columbia Metropolitan Police Department is already 38% black. Washington, D.C., is a city of 72% black. In other words, if the District of Columbia Metropolitan Police Department were to choose arbitrarily to use these scales only as the recruiting selection criteria, the supposed increase in responsive policemen would not substantially alter the black–white ratio of the force. Obviously, such a decision could never be made, in any case, since the more standard criteria for selection are needed as well as those suggested in this study.

REFERENCES

ALEX, N. *Black in blue: A study of Negro policemen.* New York: Appleton-Century-Crofts, 1969.

BOUMA, D. H. *Kids and cops.* Grand Rapids, Mich.: Ermans Publishing, 1969.

CLARK, R. *Crime in America.* New York: Simon and Schuster, 1970.

EDWARDS, G. C. *The police on the urban frontier.* New York: Institute of Human Relations Press, 1968.

HUNT, I. C., JR., & COHEN, B. *Minority recruiting in the New York City Police Department.* New York: Rand, 1971.

KELLY, R., DOCKETT, D., FARBER, M., WEST, G., JR., & GRAY, C. *The pilot police project in Washington, D.C.* Washington, D.C.: The American Institutes for Research & U.S. Government Printing Office, 1972.

KELLY, R., & WEST, G., JR. The racial transition of a police force: A profile of white and black policemen in Washington, D.C. In J. Snibbe & H. Snibbe (Eds.), *The urban policeman in transition.* Springfield, Ill.: Charles C Thomas, 1972.

MAST, R. Police–ghetto relations: Some findings and a proposal for structural change. *Race* (London: Institute for Race Relations), 1970, 11 (Whole No. 4).

RADELET, L. A. Implications of the professional concept in law enforcement for police-community relations. In H. E. Pomrenke (Ed.), *Police community relations.* Chapel Hill: University of North Carolina, Institute of Government, 1966.

STOUFFER, S. A., et al. *Measurement and prediction.* Vol. IV. *Studies in social psychology in World War II.* Princeton, N.J.: Princeton University Press, 1950.

WILSON, J. *Varieties of police behavior: The management of law and order in eight communities.* Cambridge, Mass.: Harvard University Press, 1968.

(Received April 18, 1973)

Manuscripts Accepted for Publication in the

Journal of Applied Psychology

Quality of Decisions: Individuals Versus Real and Synthetic Groups. Bertram Schoner (Department of Economics and Commerce, Simon Fraser University Burnaby 2, British Columbia, Canada), Gerald L. Rose, and G. C. Hoyt.

Effects of Numerical Data upon Understanding and Perception. William Witt (Department of Communication, University of Wisconsin, Stevens Point, Wisconsin 54481).

Influence of Dogmatism, Risk-Taking Propensity and Intelligence on Decision-Making Strategies for a Sample of Industrial Managers. Ronald N. Taylor (Faculty of Commerce and Business Administration, University of British Columbia, Vancouver 8, British Columbia, Canada) and Marvin D. Dunnette.

Interpersonal Relationships. Gerhard W. Scherf (Department of Consumer Studies University of Guelph, Guelph, Ontario, Canada).

Realistic Expectations as an Aid in Reducing Voluntary Resignations. Daniel R. Ilgen (Department of Psychological Sciences, Purdue University, West Lafayette, Indiana 47907) and William Seely.

A Performance Correlate of Personal Control Beliefs in an Organizational Context. William J. Heisler (Department of Management, University of Notre Dame, Notre Dame, Indiana 46556).

Cue Utilization and Expert Judgments: A Comparison of Independent Auditors with other Judges. Robert H. Ashton (College of Business Administration, University of Texas at Austin, Austin, Texas 78712).

Effects of Caffeine on Alertness in Simulated Automobile Driving. Gene M. Smith (Erich Lindemann Mental Health Center, Government Center, Boston, Massachusetts 02114), Edmund G. Regina, Charles G. Keiper, and Robert K. McKelvey.

Statistical Procedures for Evaluating the Practical Utility of a Moderator Approach to Prediction. Alan L. Gross (Graduate School and University Center, City University of New York, 33 West 42nd Street, New York, New York 10036), Jane Faggen-Steckler, and Karen McCarthy.

Spin-Offs From Behavioral Expectation Scale Procedures. Milton R. Blood (Department of Psychology, University of California, Berkeley, California 94720).

Brainstorming by Individuals, Coacting and Interacting Groups. Warren R. Street (Department of Psychology, Central Washington State College, Ellensburg, Washington 98926).

Mechano-Content Analysis As an Aid to Predicting Effects of Advertising Microcommunications. Donald W. Hendon (Division of Business and Economics, Columbus College, Columbus, Georgia 31907).

(*Continued on page 336*)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 264–269

EQUAL UNPLEASANTNESS MATCHING OF MALODOROUS SUBSTANCES IN THE COMMUNITY[1]

THOMAS LINDVALL[2]
Karolinska Institute, Stockholm, Sweden

LEIF T. SVENSSON[3]
Swedish Environment Protection Board, Stockholm, Sweden

Five different odorants, produced through combustion procedures of animal manure, were matched with hydrogen sulfide with regard to unpleasantness by 30 subjects. Power functions described the unpleasantness matchings for five combustion toilets tested. The exponents of the equal sensation functions indicated a relative invariance of the various combustion procedures on the exponent. Different combustion procedures resulted in different levels of odor unpleasantness, but these differences were small from a practical point of view. The experiment demonstrated the practical applicability of psychophysical methods in environmental health research.

Pollution of the atmosphere is a major concern of all nations. In particular, the general environmental problem of unpleasant odors is of increasing concern. It appears that information regarding the perceived intensity and unpleasantness of odor-producing stimuli may assist environmental planners in their efforts to reduce offensive pollutants.

Certain psychophysical methods should have application to this problem. Thus, cross-modality matching of the subjective intensity of stimuli has been carried out for several pairs of sense modalities. For example, force of handgrip has been used as a matching continuum for equating the perceived intensity of different sensory continua, such as white noise, white light, warmth, and cold (Stevens, Mack, & Stevens, 1960; Stevens & Stevens, 1960). In connection with olfaction studies, finger span has been used as the matching continuum by Berglund, Berglund, and Lindvall (1971), and Ekman, Berglund, Berglund, and Lindvall (1967). Thus, the usefulness of cross-modality matching in developing "objective" yardsticks of subjective sensory impressions has been demonstrated.

A similar method used is that of "intramodality matching" which is appropriate in developing matching scales within one sense modality. In this technique the subject matches the subjective intensity or some other attribute of, for example, two different auditory frequencies or two different odorants. That is, in cross-modality matching the matching is performed between continua, while in intramodality matching the matching is between two qualities within one modality. While intramodality intensity matching has been shown to be a reliable and valid method for vision and audition, because of technical difficulties, it has been applied only occasionally to the development of scales of odor intensity (Beck, Kruger, & Calabresi, 1954; Cain, 1969; Kruger, Feldzamen, & Miles, 1955). However, suitable olfactometers are now available for this purpose (e.g., Lindvall, 1970).

The present study is concerned with the evaluation of unpleasantness of malodorous effluents similar to those emitted from combustion toilets in recreation areas. These effluents are of growing importance as environmental nuisances. However, existing United States' guidelines for maximum permissible odor discharge from gas-fired incinerators (USA Standard, 1966) are considered unreliable and not applicable to the problem.

Specifically, this study (*a*) develops and evaluates a new measurement technique for using psychophysical methods and (*b*) investigates whether a functional relationship between physical continua found in cross-modality and intramodality matching of perceived intensity can be generalized to intramodality matching of olfactory unpleasantness.

---

[1] The investigation was supported by grants from the Swedish Environmental Protection Board and the Folksam Research Board.

[2] The authors wish to thank P. O. Olsson, S. Pettersson, and G. Rosén for their assistance with the experiment, and Trygg Engen for his valuable comments on the manuscript.

Requests for reprints should be sent to Thomas Lindvall, Department of Environmental Hygiene, Karolinska Institute, S-104 01 Stockholm, Sweden.

[3] Now at the University of Stockholm.

METHOD

Derivation of the Functional Relationship

Assuming that the power function, $\psi = a \cdot \phi^n$, adequately describes the relation between perceptual (ψ) and physical (ϕ) continua for the perceptual attributes of intensity and unpleasantness, we may write the following for two different continua, neglecting constants:

$$\psi_1 = \phi_1^m$$

$$\psi_2 = \phi_2^n.$$

In cross-modality as well as in intramodality matching the subjects are instructed to equate ψ_1 and ψ_2 at different stimulation levels. That is, the equal sensation function will have the form:

$$\phi_1^m = \phi_2^n. \qquad [1]$$

Solving for ϕ_1 and ϕ_2 in Equation 1, we obtain:

$$\phi_1 = \phi_2^{b_{12}}, \quad \text{where } b_{12} = n/m \qquad [2]$$

$$\phi_2 = \phi_1^{b_{21}}, \quad \text{where } b_{21} = m/n \qquad [3]$$

Equation 2 is the predicted function that has been experimentally obtained by cross-modality matching of the subjective intensity of stimuli on two different continua for several pairs of modalities (e.g., Stevens, 1960). This equation has also been verified experimentally for auditory signals (Stevens, 1966), for heterochromatic matching of saturation to saturation (Indow & Stevens, 1966), for vibrotactile stimuli (Stevens, 1968), and for matching of linear aliphatic alcohols (Cain, 1969). It is predicted that this equation will generalize to intramodality matching of olfactory unpleasantness.

Experimental Equipment

The laboratory used in the present study was a mobile unit employed in the sensory evaluation of odorous air pollutants. The olfactometer was based on the use of a well-defined dynamic flow of gas that permitted stable exposure concentrations, rapid changes in concentration, and fairly natural respiratory conditions (Lindvall, 1970). The equipment consisted essentially of an air-conditioned test chamber in which two teflon-coated exposure hoods had been installed. The odorant gas was mixed with purified air in fixed dilutions and added to the flow of air through the hoods by means of electromagnetically operated steel and teflon cannulae of varying bore. The unit had a special waiting room for the subjects with ventilation comparable to the actual test procedure in order to provide optimal adaptation to the experimental situation. By using suitable filter equipment (active charcoal filters and particle filters) it was possible to carry out evaluations with minimum contamination of the laboratory air.

Each odor hood received a continuous flow of 100 liters per minute of purified air corresponding to a rate of approximately .1 meters per second. A signal lamp indicated the beginning and the end of a presentation. The response panel consisted of an adjuster that could be rotated in 13 discrete steps.

Stimuli

Five types of odorous gases were produced through combustion of animal manure. The different odorants were obtained from three electrical combustion units (A, D, and E), one gas-fed furnace (B), and one oil-fed combustion unit (C).

The manure used was first homogenized and then split into samples of about .35 liter. All samples were immediately deep frozen and kept at −20°C until 24 hours before use in the combustion toilets. The dry weight of the manure was 11.9%; the content of total sulphur was .79% of the dry weight. During the experiment, the manure was burnt in the toilets while the resulting gases were fed via pipes into the olfactometer in the odor laboratory.

Absolute odor threshold determinations were carried out for all the odorous gases produced (for a description of the methods, see Lindvall, 1970). The absolute thresholds obtained were approximately the same for all combustion processes investigated (Table 1).

Each of the five combustion odors was diluted with purified and conditioned air into four steps of physical intensity in a geometrical series. Concentration values are expressed as arithmetic means of dilution factors in log parts per million (volume per volume). Stimuli are expressed as the differences between stimulus concentration values and threshold concentration values. Table 1 presents these stimulus values as well as the absolute threshold values obtained for each combustion unit. It should be noted that the stimulus values are of lower magnitude than the threshold values simply because the former are calculated as a difference between the stimulus and threshold concentration values.

Standard Odorant

It was decided that a standard odorant would be used as a measurement tool to help determine the acceptability of the complex malodorous effluents. To obtain the psychological scales for the different air pollutants, the standard odorant has to be scaled; for example, by magnitude estimation. Scale transformations can then be performed in order to find the psychological scales for differen air pollutants.

Hydrogen sulfide was chosen as the matching ga because of its practical applicability as a possibl index substance in field studies as well as its unpleasant smell. The hydrogen sulfide (Matheson

TABLE 1

ABSOLUTE THRESHOLD CONCENTRATIONS AND RELATIVE STIMULUS CONCENTRATIONS FOR THE FIVE ODORANTS

Combustion procedure	Threshold concentration	Stimulus			
Electrical combustion unit (A)	1.99	.19	.68	1.14	1.69
Gas-fed furnace (B)	2.29	.24	.72	1.23	1.82
Oil-fed combustion unit (C)	2.02	.18	.71	1.18	1.73
Electrical combustion unit (D)	2.43	.19	.65	1.21	1.76
Electrical combustion unit (E)	2.51	.19	.59	1.18	1.73

technical quality) was diluted into 13 steps of physical intensity in a geometric progression.

Chemical Analysis

Analysis of the concentration of hydrogen sulfide was carried out six to eight times a day by employing the gel technique described by Broström and Brosset (1969). The method involves absorption of hydrogen sulfide on a silver salt impregnated gel in a tube thereby coloring part of the gel. The actual exposure concentrations in the hood itself could not be analyzed due to the relative intensitivity of chemical methods available. Samples were therefore taken via a teflon tube immediately before the capillaries delivered the appropriate dosage. The concentration of hydrogen sulfide was estimated by direct measurement of the black colored part of the gel tube. Its length was compared with a standard curve based on the desorption of the compound, as obtained by a wet chemical analysis (Jacobs, Braverman, & Hochheiser, 1957).

The composition and concentration of the gas is unknown and cannot be easily chemically analyzed. Therefore the combustion odors were simply controlled by the dilution system of the olfactometer, that is, by known flows of odorous gases into the main air flow.

Subjects

Thirty female subjects, aged 20–47, participated in the experiment. They were housewives and students without any previous experience in olfaction experiments.

Procedure

The *method of intramodality magnitude matching* was used. In the left exposure hood the subject was exposed to fixed concentrations of the combustion odors, one at a time, and in the right hood to various concentrations of hydrogen sulfide adjustable in discrete steps by the subject. By rotating the adjuster, different calibrated steel and teflon cannulae were operated in or out of the main air stream to the exposure hood. The subject was asked to operate the adjuster until she found the different odors in the two hoods subjectively equal with regard to perceived unpleasantness.

In each session the subject matched hydrogen sulfide to the four different concentrations of the combustion gases. Consequently, the order of these four concentrations could be varied in 24 different ways. These possible permutations were rotated randomly over the sessions of the experiment. The experiment lasted for nine days with the subjects participating in 20-minute sessions in an irregular order. All odors were judged once a day by each subject.

In order to control adaptation and other temporal effects, as well as position effects, the matchings were performed in the following way: (*a*) The subjects had to wait for at least 15 minutes in the air-conditioned waiting room without smoking or eating. (*b*) Each subject was carefully instructed and trained in her specific task in the test chamber. (*c*) The subject had to take just one sniff in each hood: first in the hood where the combustion gases were presented, and then, immediately after exhaling, one sniff in the matching hood. If she did not perceive the odors equally unpleasant, she had to adjust the hydrogen sulfide concentration and wait for 15 seconds until a light signal indicated stable concentrations in the hoods. Then she had to make another comparison, smelling in the hoods in the same order as above. This cycle was repeated until she found the specific concentration of the combustion gas and the adjusted hydrogen sulfide concentration to be subjectively equal with respect to unpleasantness. Her decision was given to the experimenter through a signal system, and the required concentration of hydrogen sulfide was registered.

RESULTS AND DISCUSSION

The results from the 30 subjects were averaged (arithmetic means) into group data. The five equal sensation functions of unpleasantness are plotted in Figure 1. Concentrations of the matching gas (hydrogen sulfide) are read on the ordinate and relative concentrations of the combustion gases on the abscissa. Absolute combustion gas concentrations can be calculated from Table 1. The hydrogen sulfide concentrations are expressed as log parts per billion and the combustion gas concentrations in relative log units (dilution factors). From Figure 1, it is possible to determine the physical value of the intensity of hydrogen sulfide which has the same degree of unpleasantness as a given physical intensity of a certain combustion gas. A line is drawn, parallel to the ordinate, from the point on the abscissa representing the combustion concentration until it intersects the equal sensation function of the specific combustion gas. Then, another line is drawn, parallel to the abscissa, from the intersection point to the ordinate. The ordinate intersection value is the concentration of hydrogen sulfide which is considered as unpleasant as the concentration of the combustion gas. Every point in the diagram represents the average of 20 observations except for combustion gas C. In this case, each point is based on only 10 observations, owing to technical problems.

From Figure 1, it is evident that the trend of the equal sensation plots can be represented approximately by straight lines in a log-log plot. However, the standard deviation of each test point is considerable. The apparent linearity in log-log plots indicates an equal sensation function of the basic form presented in Equation 2.

In the present context the following form of the equal sensation function was judged most descriptive.

$$\frac{S_m}{S_{om}} = a \cdot \left(\frac{S_r}{S_{or}}\right)^b, \qquad [4]$$

where S_m is hydrogen sulfide concentration in parts per billion, and S_{om} the corresponding absolute threshold in the same unit.[4] S_r is combustion gas concentration in parts per million, and S_{or} is the corresponding threshold in the same unit. With the chosen form of the equal sensation function the multiplicative constant a is independent of unit of measurement for the reference and matching stimuli.

In logarithmic form Equation 3 reads:

$$\log S_m - \log S_{om} = \log a + b\,(\log S_r - \log S_{or}). \qquad [5]$$

The reference stimuli used in the experiment, $\log S_r - \log S_{or}$, can be found in Table 1 and the exponent b and the multiplicative constant a in Table 2. The obtained exponents are of the same order of magnitude for all the combustion gases studied, that is, approximately .5.

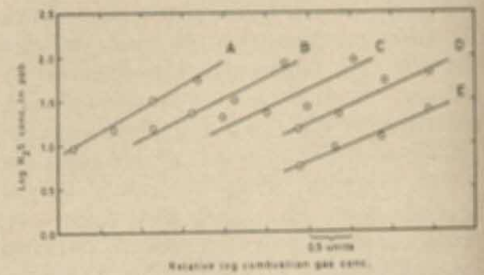


Figure 1. Five equal sensation functions in log-log coordinates obtained by matching hydrogen sulfide to five different odorants with regard to perceived unpleasantness. (The straight lines are fitted by the method of least squares and represent power functions in linear coordinates.) A, D, and E = electrical combustion units; B = gas-fed furnace; C = oil-fed furnace.

The results showed that to obtain a change in perceived unpleasantness of the same order of magnitude for both hydrogen sulfide and any one of the combustion gases, the change in log physical concentration for hydrogen sulfide had to be approximately half the change of any combustion gas ($b \sim .5$). The form of the equal sensation function for the attribute of unpleasantness within olfaction is in accordance with earlier matching studies both within and between modalities (Stevens, 1960; Stevens, 1966; Stevens, 1968; Indow & Stevens, 1966).

TABLE 2

Exponents and Multiplicative Constants of Power Functions

Combustion procedure	Exponent b	Multiplicative constant	
		a	$\log a$
Electrical combustion unit (A)	.54	5.36	.729
Gas-fed furnace (B)	.47	7.57	.879
Oil-fed combustion unit (C)	.44	8.95	.952
Electrical combustion unit (D)	.43	9.18	.963
Electrical combustion unit (E)	.39	3.42	.534

---

[4] The absolute threshold for hydrogen sulfide, .14 log parts per billion, was estimated in an earlier experiment using the same equipment (Lindvall, unpublished material).

In addition, the results clearly showed that the effect of the combustion units differ in that the same unpleasantness was caused by different concentrations for the different combustion units. For example, an unpleasantness percept corresponding to 1.49 log parts per billion hydrogen sulfide was caused by the concentrations: 2.13, 3.29, 2.93, 3.34, 4.61, measured in log parts per million, from the combustion gases A, B, C, D, and E, respectively. Although the investigation showed that combustion unit E was slightly superior to the other combustion units, the differences were so small that they could be ignored for practical purposes. It may therefore be stated that all investigated units were perceived equally unpleasant on liminal as well as supraliminal levels.

In the present context it is important to consider the subjective range of the attribute used, as illustrated by Engen and McBurney (1964). They showed that the subjective range for pleasantness of odors obtained with the method of magnitude estimation far exceeded the range of subjective intensity for different odorant concentrations estimated by the same method. In this study, the attribute unpleasantness has been used as a dimension on a quantitative rather than a qualitative continuum. Naturally, a possible standard odorant for general use in environmental health studies has to be scaled for different common attributes, for example, intensity and unpleasantness. Due consideration must be given to the choice of attribute in scaling experiments. This was convincingly shown in a study on aircraft noise where subjects were found to be able to differentiate and quantify various psychological aspects of noise, such as loudness, noisiness and annoyance (Berglund, Berglund, & Lindvall, 1974). Once these dimensional problems are solved, matchings of malodorous pollutants to a standard odor will give the psychological scales for the pollutants simply by scale transformations.

Need for a Standard Odor

From the point of view of basic research, it is desirable to make the relative values obtained in olfactory research comparable from one laboratory to another by including a standard odorant in all scaling experiments. A similar view was expressed by Cain (1971) who also advocated validation of the exponent in the equal sensation function by the use of the same experimental setting with the same standard odor and with other odors being scaled individually and matched by pairs.

Undeniably, there are a lot of basic psychological problems to be solved before one has achieved a valid and reliable relative measure of unpleasantness against a standard odorant. However, the present experiment shows the applicability of the method of equal sensation matching in environmental health studies of malodors. In spite of the complexity of the malodorous gases and the simple physical measures used for both the standard gas and the combustion gases, the functional relationships obtained are simple and congruent with earlier findings in the area.

To be of general value the psychophysical method presented in this study has to be validated in a large-scale experiment with several odors, each being used as the fixed and variable gas. For use in field studies, for example, when comparing different malodorous emission sources, the standard odor used for matching must be highly reproducible technically and must have a wide psychological range. In addition, it would be desirable if this standard yielded a simple relation between attributes such as unpleasantness and intensity. Hydrogen sulfide can be suggested as such a standard odor.

Practical Implications

The results of the present investigation demonstrated the practical applicability of unpleasantness matching in environmental psychology. It was thus shown that unpleasantness evoked by malodorous combustion gases of animal manure was independent of the combustion procedure tested. From a perceptual point of view the combustion toilets were judged equally efficient. The unpleasantness levels for undiluted combustion gases corresponded to approximately 1 part per million hydrogen sulfide which was perceived as rather unpleasant and comparable to the smell of rotten eggs. (The threshold for hydrogen sulfide was approximately 1/1,000 parts per million.)

In addition, with the knowledge of the parameters of the psychophysical function in connection with atmospheric dispersion calculations, it was assessed that houses with such toilets or recreation areas must be at least 200 meters apart in order to reduce the detectability of the odor to an acceptable degree.

REFERENCES

BECK, L. H., KRUGER, L., & CALABRESI, P. Observations of olfactory intensity. I. Training procedure, methods, and data for two aliphatic homologous series. *Annals of the New York Academy of Science*, 1954, [illegible], 225–238.

BERGLUND, B., BERGLUND, U., & LINDVALL, T. On the principle of odor interaction. *Acta Psychologica*, 1971, 35, 255–268.

BERGLUND, B., BERGLUND, U., & LINDVALL, T. Scaling loudness, noisiness, and annoyance of aircraft noise. *Reports from the Psychological Laboratories, University of Stockholm*, 1974, No. 393.

BROSTRÖM, C. [illegible], & BROSSET, C. A method for simultaneous determination of hydrogen sulfide and sulfur dioxide in flue gases. *Atmospheric Environment*, 1969, 3, 407–416.

CAIN, W. S. Odor intensity: Differences in the exponent of the psychophysical function. *Perception & Psychophysics*, 1969, **6**, 349–354.

CAIN, W. S. Physicochemical characteristics and supraliminal odor intensity: Reply to Mitchell. *Perception & Psychophysics*, 1971, **9**, 478–479.

EKMAN, G., BERGLUND, B., BERGLUND, U., & LINDVALL, T. Perceived intensity of odor as a function of time of adaptation. *Scandinavian Journal of Psychology*, 1967, **8**, 177–186.

ENGEN, T., & MCBURNEY, D. H. Magnitude and category scales of the pleasantness of odors. *Journal of Experimental Psychology*, 1964, **68**, 435–440.

INDOW, T., & STEVENS, S. S. Scaling of saturation and hue. *Perception & Psychophysics*, 1966, **1**, 253–271.

JACOBS, M. B., BRAVERMAN, M. M., & HOCHHEISER, S. Ultramicrodetermination of sulphides in air. *Analytical Chemistry*, 1967, **29**, 217.

KRUGER, L., FELDZAMEN, A. N., & MILES, W. R. A scale for measuring supra-threshold olfactory intensity. *American Journal of Psychology*, 1955, **68**, 386–395.

LINDVALL, T. On sensory evaluation of odorous air pollutant intensities. *Nordisk Hygienisk Tidskrift*, 1970, Suppl. 2, 1–181.

STEVENS, J. C., MACK, J. D., & STEVENS, S. S. Growth of sensation on seven continua as measured by force of handgrip. *Journal of Experimental Psychology*, 1960, **59**, 60–67.

STEVENS, J. C., & STEVENS, S. S. Warmth and cold-dynamics of sensory intensity. *Journal of Experimental Psychology*, 1960, **60**, 183–192.

STEVENS, S. S. On the new psychophysics. *Scandinavian Journal of Psychology*, 1960, **1**, 27–35.

STEVENS, S. S. Power-group transformations under glare, masking, and recruitment. *Journal of the Acoustical Society of America*, 1966, **39**, 725–735.

STEVENS, S. S. Tactile vibration: Change of exponent with frequency. *Perception & Psychophysics*, 1968, **3**, 223–228.

USA STANDARD. Approach requirements for domestic gas-fired incinerators. *American Gas Association Inc.* New York, December 28, 1966.

(Received April 9, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 270-273

EVALUATION OF A DECELERATION WARNING LIGHT FOR REDUCING REAR-END AUTOMOBILE COLLISIONS

JOHN VOEVODSKY[1]

Portola Valley, California

An amber light, center mounted on the rear of several hundred San Francisco taxicabs, was designed to communicate information about the taxi's deceleration to following drivers. The light, activated by the use of the brake pedal, was pulsed in a controlled fashion at a rate, duty cycle, and intensity that varied exponentially with a component of deceleration. A comparison of an experimental group (with the light for 12.3 million miles) with a control group (without the light for 7.2 million miles) revealed the light prevented 5.4 collisions, 1.02 cab driver injuries, and $643 of taxicab damage per million miles.

The possibility of designing automotive equipment to prevent automobile collisions has been greatly overshadowed by attempts to design automotive equipment to protect occupants during a crash (Bowen, 1972). Prevention of a significant number of rear-end collisions may be possible by the use of a light system warning other drivers that the car ahead is decelerating. The theoretical basis for the design of such a deceleration warning light system was published earlier (Voevodsky, 1967).[2] The present study is an evaluation of the collision prevention effectiveness of a deceleration warning light under normal driving conditions.

METHOD

The deceleration warning light system designed for this study was an amber warning light, center mounted on the rear of a vehicle at the same height as existing stop lights, which communicated a component of deceleration initiated by the driver of that vehicle to the driver of a following vehicle. A device for measuring deceleration, essentially an inertial pendulum, hence independent of any mechanical adjustments of the car such as brake pedal position, was rigidly attached to the vehicle. The warning light, activated by use of the brake pedal, was pulsed in a controlled fashion at a rate that varied exponentially with the component of deceleration. A 50% increase in frequency of the light pulses from 1.0 pulse per second to 7.6 pulses per second for each .1 gravity force increase in deceleration from 0 to .5 gravity provides the exponential relationship.

Subjects

A commercial taxicab company, to be referred to as the Cab Company, participated in the experiment at no cost to itself other than time out of service for equipping the cabs and operating them so equipped.[3] The Cab Company was asked to make their San Francisco fleet of 503 cabs available for two reasons: (*a*) The San Francisco fleet travels almost 2 million miles per month, thereby enabling a testing of the equipment over significant mileage in a relatively short time period. (*b*) The Cab Company's complete records of cab accidents were available for the five years prior to 1972, thus providing a statistical base from which to evaluate the test data.

Procedure

On March 1, 1972, 343 taxis began operation with deceleration warning lights. The remaining 160 taxis were to be traded in on new vehicles and therefore were not equipped, but served as a concurrent control group. The new vehicles were equipped when delivered to the Cab Company. Later, between June 1 and August 28, 1972, the deceleration warning lights were removed from previously equipped cabs in order to equalize the mileage driven by the equipped and nonequipped cabs. This approximately equal monthly mileage was continued through the end of the experiment, February 1, 1973. The changes in the proportions of cabs equipped or unequipped were corrected by expressing all collision rates as collisions per million miles driven.

The dispatcher assigned taxi drivers to equipped or unequipped taxis at his discretion, disregarding any preferences expressed by the drivers.

No massive education program on the operation and function of the deceleration warning light system was undertaken prior to the experiment or

---

[1] Request for reprints should be sent to John Voevodsky, 177 Goya Road, Portola Valley, California 94025.

[2] This report is not concerned with the engineering problems encountered while developing the equipment.

[3] Acknowledgment is made of the cooperation of the Yellow Cab Company of San Francisco for their participation and for making their accident records available.

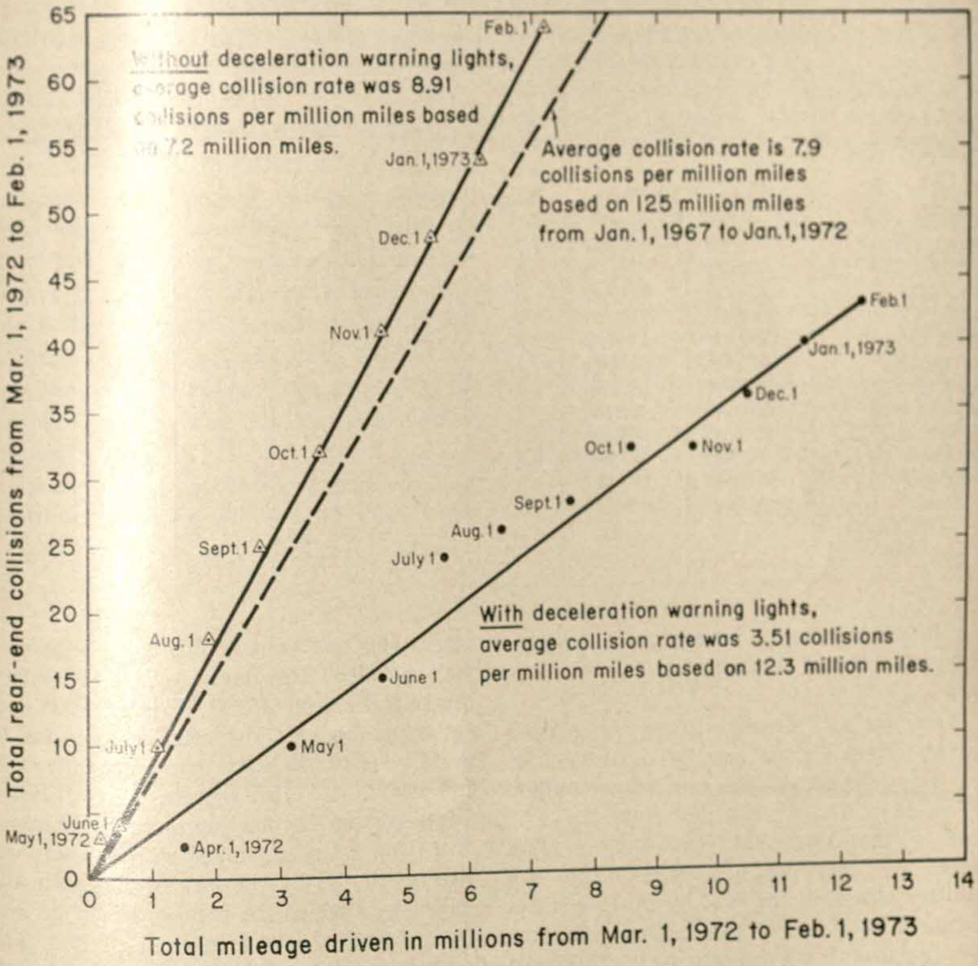


FIGURE 1. Rear-end collisions of cabs equipped and not equipped with deceleration warning light.

during the experiment. Limited news coverage at the start of the test provided the only source of public information.

RESULTS

At the end of the reporting period, January 31, 1973, the vehicles equipped with the deceleration warning lights had traveled a total of 12.3 million miles with a rear-end collision rate of 3.51 collisions per million miles (Figure 1).[4] This is a 60.6% reduction in the rear-end collision rate from the concurrent control group which had 8.91 collisions per million miles over a total of 7.2 million miles. The rate for the concurrent control group is almost identical to the 1971 rate of 8.89 over 21.5 million miles, and is near the mean rate of 7.9 for the five years prior to January 1, 1972, computed over 125 million miles from the Cab Company records.

The 10 months for which comparative data are available can be taken as 10 independent experiments. The month of April 1972 was excluded from the analysis as there were no control data for this month. When the rear-end collisions are treated as collisions per million miles within each month, a paired *t*

---

[4] The months appear at different points above the abscissa for equipped and nonequipped cabs because in the earlier months the total mileage driven by the equipped cabs was greater than that of the nonequipped cabs.

TABLE 1

REAR-END COLLISIONS, DRIVERS INJURED, AND CAB DAMAGE AS A FUNCTION OF MILES DRIVEN FOR CABS EQUIPPED AND NOT EQUIPPED WITH DECELERATION WARNING LIGHTS

Month/year	Mileage		Rear-end collisions		Drivers injured		Cab damage (dollars)	
	With	Without	With	Without	With	Without	With	Without
March 1972	1,515,793	197,903	2	3	0	1	[illegible]	1,275
April 1972	1,661,455	0	8	0	2	0	[illegible]	0
May 1972	1,439,438	245,316	5	1	1	0	[illegible]	0
June 1972	987,941	711,234	9	6	1	1	[illegible]	227
July 1972	959,977	757,510	2	8	0	0	[illegible]	1,408
August 1972	1,085,868	826,081	2	7	0	0	[illegible]	425
September 1972	921,800	855,596	4	7	0	1	[illegible]	765
October 1972	1,038,207	1,002,130	0	9	0	2	[illegible]	2,124
November 1972	876,549	844,301	4	7	3	1	[illegible]	189
December 1972	875,895	835,591	4	6	0	4	[illegible]	853
January 1973	899,105	908,877	3	10	1	2	[illegible]	211
Totals	12,262,028	7,184,539	43	64	8	12	[illegible]	7,477
Per Million Miles			3.51	8.91	.65	1.67	[illegible]	1,041

test of the differences for equipped and nonequipped cabs is significant ($t = 4.06$, $df = 8$, $p = .006$).[5]

The rate of personal injury to taxicab drivers struck in the rear by other vehicles for the equipped vehicles was .65 per million miles, or a 61.1% reduction from the 1.67 rate of the concurrent control group (see Table 1). The costs of repairs to the cab vehicles struck in the rear by other vehicles declined from $1,041 per million miles for the concurrent control group to $398 per million miles for the deceleration warning light vehicles. This is a reduction of 61.8%. Damage figures for the vehicles that collided with taxis are not known. Because the Cab Company is not at fault these amounts are not included in their records.

DISCUSSION

A common criticism of social experiments that involve innovative arrangements is that the operators are sensitive to the new arrangements and behave differently because of the so-called Hawthorne effect (Roethlisberger & Dickson, 1939). This criticism does not apply because the collisions are due to the drivers of the cars that follow the taxis and not to the taxi drivers. That the taxi driver has not become more cautious because of driving a taxi equipped with a safety device is attested by the lack of change in the collisions in which the taxicab driver ran into the rear-end of a car in front of him. The drivers of equipped taxis produced 7.1 such collisions per million miles, and those not equipped produced 8.1 such collisions per million miles. When this difference is tested statistically by the same procedure previously described, the difference is not significant ($t = .65$, $df = 8$, $p = .52$).

The results of this 11-month, 19.5-million mile test of a deceleration warning light system in natural driving conditions indicate that a significant reduction in rear-end collisions and injury can be achieved. During the experiment, the deceleration warning light prevented rear-end collisions to equipped vehicles by 5.4 collisions per million miles, reduced the number of cab drivers injured from such collisions by 1.02 injured per million miles, and reduced the cost of repairs to cab vehicles struck in the rear by $643 per million miles. These figures do not include taxi passenger injuries and the personal injuries and property damage prevented for the vehicles which ran into the rear of the cabs. With the United States public driving approximately 1,115,000 million miles each year (National Safety Council, 1971) the potential reduction in personal injury and property damage from rear-end collisions for American society is enormous. The same report from the National Safety Council indicates that rear-end collisions account for 2,000 deaths per year. If [illegible]0% of these are avoidable, the savings of [illegible],000 lives per year would be a human benefit.

---

[5] The division of accident rates into units of months is somewhat arbitrary, but to produce a statistical test at all it was necessary to use the data available from the Cab Company. Mileage data are reported monthly, rather than by miles to accident, which would permit the use of Poisson statistics.

Hence, this quantitative experiment also illustrates the kind of "on the street" field evaluation possible with the cooperation of public and private agencies.[6]

---

[6] The California Highway Patrol granted the permit to place the deceleration warning lights on the cabs for the experiment. This investigation could not have been possible without their approval.

REFERENCES

BOWEN, W. Auto safety needs a new roadmap. *Fortune*, April 1972, **85**, 98–101; 142–145.

NATIONAL SAFETY COUNCIL. *Accident facts*. Chicago: Author, 1971.

ROETHLISBERGER, F. J., & DICKSON, W. J. *Management and the worker*. Cambridge, Mass.: Harvard University Press, 1939.

VOEVODSKY, J. Inferences from visual perception and reaction time to requisites for a collision-preventing cyberlite stop lamp. *Proceedings of the National Academy of Sciences of the United States of America*, 1967, **57**, 688–695.

(Received May 14, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 274–280

EFFECTS OF DIFFERENT LEVELS OF INTEGRATION ON ADVERTISING PREFERENCE AND INTENTION TO PURCHASE

GEORGE J. SZYBILLO[1] AND JACOB JACOBY

Purdue University

Ninety blacks and 90 whites evaluated six proposed advertisements in terms of attractiveness and likelihood of purchase of the product advertised. The advertisements differed in the degree of racial integration depicted, which was operationally defined by a 5-point continuum ranging from 0 whites and 4 blacks (0W-4B) to 4 whites and 0 blacks (4W-0B). The hypothesis that young blacks react negatively to integrated ads depicting tokenism (e.g., three whites to one black) and positively to ads depicting equality (e.g., two whites to two blacks) was supported with each of the dependent measures.

Bauer and Cunningham (1970) have noted that "the ways in which blacks buy and use goods and services are simultaneously a practical business concern and a social issue, with the two interacting with each other [p. 3]." The manufacturer wishing to effectively segment this market and simultaneously maintain his white market may face a number of social problems and dilemmas related to such things as product design, brand perceptions, labeling, merchandising, and especially advertising.

With respect to advertising, it might be asked whether black consumers and white consumers have different preferences for all black, all white, or integrated advertising? Is there a possibility of backlash from whites or, perhaps more interesting, from blacks viewing integrated advertising? Does integrated advertising have any differential effect on the purchase intentions of blacks and whites? Moreover, what constitutes an "integrated" advertisement? Is it the appearance of only 1 black in a group of 30?

By and large, evidence adduced thus far (Barban, 1969; Barban & Cundiff, 1964; Barban & Grunbaum, 1965; Guest, 1970; Stafford, Birdwell, & Van Tassel, 1970; Tolley & Goett, 1971) suggests that white subjects display about the same degree of positive affect to integrated advertising as they do to all white ads (i.e., ads containing all white characters), whereas black subjects tend to prefer integrated or all black advertising over all white ads. Notable exceptions to this generalization are Cagley and Cardozo's (1970) finding that highly prejudiced white subjects reacted negatively to integrated advertisements, and Gould, Sigband, and Zoerner's (1970) finding that younger (i.e., 14–29 years of age) blacks were negatively predisposed to the concept of integrated advertising.

While some controversy exists regarding the use of black models in advertisements (Kassarjian, 1969; Wheatley, 1971), Kassarjian (1971) contended that the black model is still only a token black. Integrated ads typically feature a single model smoking, drinking, modeling clothes, etc. together with a number of white models. Wall (1970) suggested that although the intent of the advertiser is to establish empathy, blacks respond primarily with a sense of loneliness and alienation and become uncomfortable.

This raises the entire issue of just what constitutes an integrated advertisement. Boyenton (1965) defined integrated advertising as the inclusion of blacks in general consumer advertising and indicated that the same aspect which characterizes blacks among the general population characterizes blacks in an advertisement, namely, "the dark face." However, such a definition does not reflect the fact that there can be varying degrees of integration depicted in an ad. Conceivably, responses to an integrated advertisement will be different when the advertisement contains one black model and five white models versus when it contains one white model and five black models. The few studies which have examined the reaction of blacks and whites to more than one level of integration have collapsed the levels into a single level called "integration," without providing data indicating no significant differences across the different levels of integration (Barban, 1969; Stafford et al., 1970). Thus, none of the studies conducted so far has explicitly incorporated more than a single level of "integrated advertising" into its experimental design.

---

[1] The authors wish to thank Barry Barnes and Diane Saint for their assistance in collecting the data for this experiment.

Requests for reprints should be sent to George J. Szybillo, Graduate School of Business Administration, New York University, 100 Trinity Place, New York, New York 10006.

Accordingly the present experiment was designed to examine the reactions of young blacks and whites to proposed advertisements composed of different ratios of black principal characters and white principal characters. Based on earlier findings, it was hypothesized that young white respondents react equivalently, both in terms of affect and purchase intentions, across all levels of all white and integrated advertising. Also, based on the earlier literature, particularly the findings of Gould et al. (1970) and the writing of Wall (1970), it was hypothesized that young black respondents react negatively to an advertisement whose racial composition suggests tokenism and favorably to advertisements whose racial composition depicts black–white equality.

Since Gould et al. (1970) found negative affect directed toward the concept of "integrated advertising" only with blacks below age 29, this last hypothesis is most properly tested using young blacks. Other reasons for using college-age black subjects and white subjects included the following: (*a*) Since Barban's early studies (Barban & Cundiff, 1964; Barban & Grunbaum, 1965) no reports have been published that compare black student and white student reactions to integrated advertising. As social conditions change, so too should there be a measurable impact on people's perception of those changes. (*b*) The young black college-educated population represents a market segment with considerable economic impact (Wattenberg & Scammon, 1973) and, as such, is worthy of study in its own right. (*c*) Business firms have become more concerned with how they can attract minority employees. Integrated advertising does more than just sell the product; it can also give minority consumers impressions about the nature of a manufacturer's hiring policies (Szybillo, Jacoby, & Busato, 1973).

METHOD

Subjects

The subjects were 90 black and 90 white male volunteers who were students at Purdue University during the 1973 spring semester. Subjects were randomly assigned to experimental treatments.

Experimental Design

The experimental design was a 2 × 5 × 6 (Race of Respondents × Racial Ratio of Advertisement × Products) factorial. The first two factors were between-subjects factors, while the third was a within-subjects factor (Winer, 1971, p. 566). The five levels of the second factor were (*a*) 100% black (four blacks, no whites), (*b*) 75% black (three blacks, one white), (*c*) 50% black (two blacks, two whites), (*d*) 25% black (one black, three whites), and (*e*) 0% black (no blacks, four whites). The six products chosen for study were Irish whisky, dress shirt, umbrella, attaché case, deodorant, and underwear. In an attempt to reduce demand characteristics (cf., Orne, 1962), each subject received the same racial ratio in each of the six product advertisements to which he was exposed. As described below, an additional 15 white and 15 black subjects served as a control.

Independent Variables

The independent variables of primary interest were race of respondent and racial ratio of advertisement. Race of respondent was treated as a classification variable: black versus white. The racial ratio variable was presented to the subjects by manipulating the extent of integration in the six proposed product advertisements.

To eliminate variability due to different brand names and quality of execution associated with different ads, actual product advertisements were not used in this study. Subjects were told that a national magazine publisher was interested in obtaining ratings of descriptions of proposed advertisements. Each proposed advertisement contained three or four sentences of product descriptive copy taken from actual advertisements, an age range for the models in the ad (i.e., 20–24), the initials of a fictitious advertiser, the name of a fictitious manufacturer, and the sex, number, and race of the models in the advertisement. The 30 control subjects received the identical material, with the exception of information regarding race of the models (which was omitted).

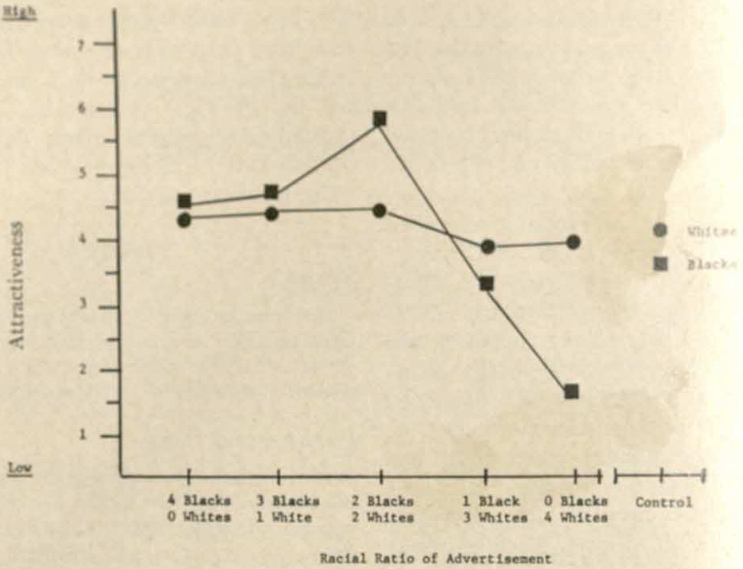


FIGURE 1. Race of Respondent × Racial Ratio of Advertisement Interaction for Analysis of Advertisement Attractiveness.

Dependent Variables

Two types of reactions were considered: (*a*) preference (the dependent variable most often used in earlier studies) and (*b*) purchase intentions. Preference was measured using three different 7-point bipolar scales. These scales were attractiveness of the advertisement (having endpoints labeled "unattractive" and "attractive"), the personal appeal of the advertisement (having endpoints labeled "does not appeal to me" and "does appeal to me"), and the overall appeal of the advertisement (having endpoints labeled "bad" and "good"). Likelihood of purchase was measured using a 7-point bipolar scale (having endpoints labeled "very unlikely to purchase" and "very likely to purchase"). The positive ends of these scales were alternated to reduce response bias. In this report, high numbers represent the positive end of the scales.

Each subject completed each of the dependent measures for proposed ads for each of the aforementioned products. The order of the advertisements was randomized in each subject's questionnaire booklet.

Procedure

The white subjects and black subjects received questionnaires in separate rooms. A white proctor administered the questionnaire to white respondents, and a black proctor administered the questionnaire to black respondents. The respective proctor remained with the subjects at all times to ensure that subjects understood directions and did not communicate with each other.

Subjects were instructed to answer all questions in their questionnaires. On completion, subjects received a postexperimental questionnaire probing their perceptions of the intent of the study. A manipulation check for the racial ratio variable was also included. Before the subjects left the testing session they were debriefed as to the genuine purpose of the experiment.

RESULTS

The subjects' responses to the manipulation check item on the postexperimental question-

TABLE 1

ANALYSES FOR ATTRACTIVENESS AND LIKELIHOOD OF PURCHASE RATINGS

Source	*df*	Attractiveness		Likelihood of purchase	
		MS	*F*	*MS*	*F*
Between subjects					
Race of respondent (A)	1	12.48	4.31*	14.19	7.95**
Racial ratio of advertisement (B)	4	143.90	49.72**	51.55	28.89**
A × B	4	86.07	29.74**	42.16	23.63**
Error (between)	140	2.89		1.78	
Within subjects					
Product (C)	5	23.89	14.89**	3.83	5.53**
A × C	5	9.47	6.10**	1.90	2.75**
A × C	20	2.59	1.67*	.90	1.30
A × B × C	20	2.17	1.40	1.05	1.51
Error (within)	700	1.55		.69	

* $p < .05$.
** $p < .01$.

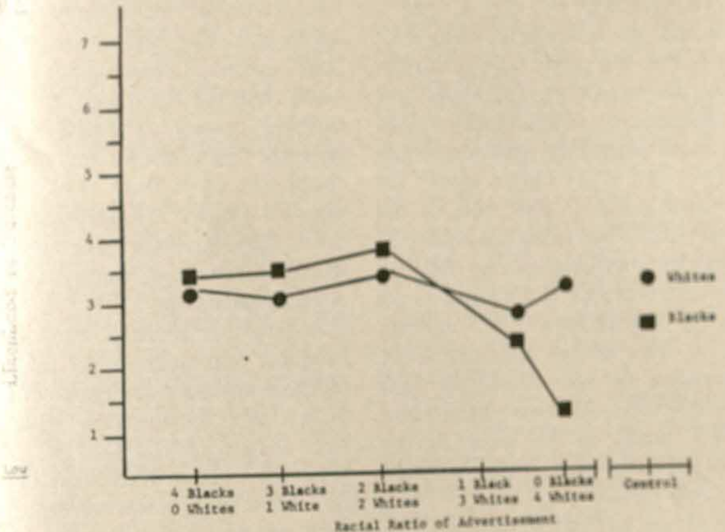


FIGURE 2. Race of Respondent × Racial Ratio of Advertisement Interaction for Analysis of Likelihood of Purchase.

naire indicated that all subjects were aware of the racial ratio which appeared for each of their product advertisements. Responses to the item probing for the presence of demand characteristics indicated that no subject guessed the purpose of the study was to determine the effect of *different* racial ratios on their responses. It appears that providing each subject with only *one* racial ratio across the six product advertisements was effective in reducing demand characteristics.

Table 1 presents the F ratios and associated probability values obtained from analyses of variance for the dependent measures. Race of respondent had a significant effect for each analysis. In addition, the main effects for racial ratio of advertisement and products were also significant for both analyses. However these main effects should be considered with respect to the significant interactions. Importantly, in terms of the purpose of this experiment, the interaction between the race of respondent factor and the racial ratio of advertisement factor was significant for both dependent measures. The significant Race of Respondent × Racial Ratio of Advertisement interactions for attractiveness and likelihood of purchase are depicted in Figures 1 and 2, respectively.

Consideration of the control condition data (cf., Figures 1 and 2) for the black subjects and white subjects provides interesting perspective. Across both dependent measures, there was virtually no difference between the ratings assigned by the whites to the all white and control ads, but there were considerable differences for the black subjects, who rated the all white ads substantially lower. This probably occurred because, in the absence of any information regarding racial composition of the ad models, most people (blacks as well as whites) would assume that the characters were all white. However, making this fact (zero blacks to four whites) salient for blacks served to lower attractiveness and intention to purchase ratings.

A Newman-Keuls analysis of the Race of Respondent × Racial Ratio of Advertisement interaction for the attractiveness of advertisement measure indicated that: (*a*) blacks did not rate the conditions of four blacks to zero whites or three blacks to one white as being more attractive than did whites; (*b*) blacks rated the condition of two blacks to two whites as more attractive than did whites ($p < .01$); and (*c*) blacks rated the conditions of one black to three whites and zero blacks to four whites as less attractive than did whites (respectively, $p < .05$, $p < .01$).

Analysis of the black subjects' mean ratings of the levels of the racial ratio of advertisement factor indicated that: (*a*) black subjects found the condition of two blacks to two whites to be more attractive than any other racial ratio ($p < .01$); (*b*) blacks found the condition of three blacks to one white or the all black model condition to be more attractive than either the conditions of one black to three whites or zero blacks to four whites ($p < .01$); and (*c*) blacks found the condition of one black to three whites more attractive than the condition of zero blacks to four white models ($p < .01$). Analysis of the white subjects' mean ratings of the levels of the racial ratio of advertisement factor revealed no significant differences.

A Newman-Keuls analysis of the Race of Respondent × Racial Ratio of Advertisement interaction for the measure of likelihood of purchase indicated that: (*a*) blacks, in comparison to whites, did not assign greater Likelihood of Purchase ratings to the condition of four blacks to zero whites or the condition of three blacks to one white; (*b*) black subjects did assign greater likelihood of purchase ratings to the condition of two blacks to two whites than did white subjects ($p < .05$); and (*c*) blacks had lower likelihood of purchase ratings than did whites when viewing either the conditions of one black to three whites or zero blacks to four whites (respectively, $p < .05$, $p < .01$).

Analysis of the black subjects' mean ratings of the levels of the racial ratio of advertisement factor indicated that the results obtained for likelihood of purchase were identical to those obtained from the analysis of attractiveness of advertisement ratings. Analysis of the white subjects' mean ratings of the levels of the racial ratio of advertisement factor revealed no significant differences.

A Newman-Keuls analysis of the Race of Respondent × Products interaction for the attractiveness measure indicated that: (*a*) whites assigned greater attractiveness ratings than did the blacks to Irish whisky, dress shirt, and attaché case (respectively, $p < .01$, $p < .01$, and $p < .05$) and (*b*) white subjects' ratings of attractiveness did not differ from those of black subjects for umbrella, deodorant, and underwear. Analysis of the black subjects' mean ratings of products indicated that the dress shirt advertisement was assigned greater attractiveness ratings than were the Irish whisky and underwear advertisements ($p < .01$). The white subjects assigned greater attractiveness ratings to the dress shirt advertisement than to the Irish whisky, underwear, umbrella, and deodorant advertisements ($p < .01$).

An analysis of the Race of Respondents × Products interaction for the likelihood of purchase measure indicated that whites were more likely to purchase Irish whisky than were blacks ($p < .05$). The black subjects indicated greater likelihood of purchase for the dress shirt advertisement than for the Irish whisky advertisement ($p < .01$), and white subjects indicated greater likelihood of purhcase for the dress shirt advertisement than for the umbrella advertisement or the deodorant advertisement ($p < .01$).

Analysis of the Racial Ratio of Advertisement × Products interaction for the attractiveness measure indicated that, regardless of product advertisement, the ratio of two blacks to two whites received higher attractiveness ratings than did either the condition of three whites to one black or the condition of zero blacks to four whites ($p < .05$). The ratio of two blacks to two whites also received higher attractiveness ratings than did the condition of four blacks to zero whites for the dress shirt, umbrella, deodorant, and underwear advertisements ($p < .05$). For two of the racial ratios (i.e., two blacks to two whites, and three blacks to one white) the dress shirt advertisement was rated as more attractive than any other product advertisement ($p < .01$). A similar product trend was observed in the other racial ratio conditions.

DISCUSSION

As predicted, young blacks rated an advertisement suggesting equality (two blacks to two whites) more favorably than they did an advertisement implying tokenism, and they even found the former advertisement more attractive than did white subjects. Whites, on the other hand, rated the advertisement indicative of black tokenism (one black to three whites) as more attractive than did blacks. This last finding is the reverse of that found by Barban and Cundiff (1964) who used a comparably integrated advertisement (i.e., one black to two whites).

Both black subjects and white subjects of this study showed the closest agreement in assigning attractiveness ratings to the all black model condition, less agreement on the integrated conditions, and the least agreement on the all white condition. In contrast, Barban (1969) found that black subject and white subject agreement was greatest for all white ads, less so for integrated ads, and least so for all black ads. The results obtained in this study may be a function of (*a*) cultural changes over the years, (*b*) the change in characteristics of the subject populations of the two studies (Barban's subjects were not college students, but were middle-income blacks and whites.), or (*c*) the different type of advertising stimuli used in the studies. Similar to other findings of Barban (1969), blacks in the present study found integrated or all black advertisements to be more attractive than all white advertisements. The results are in accord with the previous findings of Stafford et al. (1970), Guest (1970), and Tolley and Goett (1971); no white backlash, as measured by attractiveness, was found for any of the racial ratios depicting integration.

The results of the analysis of variance on likelihood of purchase are interesting. Black subjects showed greater likelihood of purchase than did whites for the ads of three blacks to one white and two blacks to two whites. However, they showed less likelihood of purchase than did whites when viewing the advertisements of one black to three whites or the all white model advertisements. Blacks showed greater likelihood of purchase to all black model ads or any form of integrated ad in comparison to the all white model ads. Whites, surprisingly, showed greater likelihood of purchase when the advertisement was either all black, all white, or equally integrated than when there were three white models and one black model in the ad. This may indicate that college age whites also react negatively to tokenism in advertisements.

The results presented in this study are based on descriptions of "proposed" advertisements. In order to enhance the internal validity (Campbell & Stanley, 1963) of the study, confounding factors such as position of the models, attractiveness of the models, and roles of the models were not included in the ad descriptions. These factors may effect the impact of racial ratio when judging the external validity of these results; however, this remains to be empirically verified. Additionally further research is necessary with different black subject and white subject populations.

Probably the most significant finding of this investigation is that different levels of integration are capable of generating different responses from members of the same racial group. Since integration does vary along a continuum, subsequent research in the area should either employ different levels of racial composition of an ad as a factor in the experimental design or, at very least, take cognizance of this fact.

REFERENCES

BARBAN, A. M. The dilemma of integrated advertising. *Journal of Business*, 1969, **42**, 447–496.

BARBAN, A. M., & CUNDIFF, E. W. Negro and white response to advertising stimuli. *Journal of Marketing Research*, 1964, **1**, 53–56.

BARBAN, A. M., & GRUNBAUM, W. F. A factor-analytic study of Negro and white responses to advertising stimuli. *Journal of Applied Psychology*, 1965, **49**, 274–279.

BAUER, R. B., & CUNNINGHAM, S. M. The Negro market. *Journal of Advertising Research*, 1970, **10**, 3–13.

BOYENTON, W. F. The Negro turns to advertising. *Journalism Quarterly*, 1965, **42**, 227–235.

CAGLEY, J. W., & CARDOZO, R. N. White response to integrated advertising. *Journal of Advertising Research*, 1970, **10**, 35–39.

CAMPBELL, D. T., & STANLEY, J. C. *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally, 1963.

DOMINICK, J. R., & GREENBERG, B. S. Three seasons of blacks on television. *Journal of Advertising Research*, 1970, **10**, 21–27.

GOULD, J. W., SIGBAND, N. B., & ZOERNER, A. E. Black consumer reactions to integrated advertising: An exploratory study. *Journal of Marketing*, 1970, **34**, 20–26.

GUEST, L. How Negro models affect company image. *Journal of Advertising Research*, 1970, **10**, 29–33.

KASSARJIAN, H. H. The Negro and American advertising, 1946–1965. *Journal of Marketing Research*, 1969, **6**, 29–39.

KASSARJIAN, H. H. Blacks in advertising: A further comment. *Journal of Marketing Research*, 1971, **8**, 392–393.

ORNE, M. T. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their complications. *American Psychologist*, 1962, **17**, 776–783.

STAFFORD, J. E., BIRDWELL, A. E., & VAN TASSEL, C. E. Integrated advertising—White backlash? *Journal of Advertising Research*, 1970, **10**, 15–20.

SZYBILLO, G. J., JACOBY, J., & BUSATO, J. Effects of integrated advertising on perceived corporate hiring policy. *Proceedings of the 81st Annual Convention of the American Psychological Association*, 1973, **8**, 819–820. (Summary)

TOLLEY, B. S., & GOETT, J. J. Reactions to blacks in newspaper ads. *Journal of Advertising Research*, 1971, **11**, 11–17.

WALL, K. A. The great waste: Ignoring blacks. *Marketing/Communications*, [illegible], 42–50.

WATTENBERG, B. J., & SCAMMON, [illegible] Black progress and liberal rhetoric. *Commentary*, 1973, **55**, 35–44.

WHEATLEY, J. J. The use of black models in advertising. *Journal of Marketing Research*, 1971, **8**, 390–392.

WINER, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1971.

(Received May 7, [illegible])

Journal of Applied Psychology
1974, Vol. 59, No. 3, 281-286

A DYNAMIC MODEL OF RELATIONSHIPS AMONG ADVERTISING, CONSUMER AWARENESS, ATTITUDES, AND BEHAVIOR

DAVID A. AAKER[1]
University of California, Berkeley

GEORGE S. DAY
Stanford University

Nineteen telephone surveys, made at two-month intervals to a national probability sample of 1,200 households who were users of instant coffee, yielded aggregate measures of brand awareness, attitudes towards brands, and advertising exposure. These measures and an aggregate purchase behavior measure were used to estimate three multiple regression equations that formed a dynamic recursive model of the relationships between the measures. Advertising influenced both awareness and attitude, while awareness and attitude influenced behavior. The influence of advertising went directly from awareness to behavior, not through attitude. Thus, the hierarchy model of communication effects was found to be only partially correct. Attitude change preceded behavior change, although a less significant reverse effect also emerged.

One of the most durable and widely adopted models of the communication process is the notion of a hierarchy of effects that postulates a causal flow from cognitive change to attitude change to behavioral change. In the marketing context this model hypothesizes that advertising has its most direct and discernable effect on cognitions, a lesser effect on attitudes, and influences behavior primarily through its impact on cognitions and attitudes. (Lavidge & Steiner, 1961; McGuire, 1968) This formulation has been sharply attacked on two grounds: First, it is argued that all individuals do not necessarily proceed through each stage. Second, the direction of the flow has been challenged—in particular, there may be situations in which attitude change follows behavioral change (Kiesler, Collins, & Miller, 1970; Krugman, 1965).

The purpose of this article is to examine the communication process, using a large scale field test, with emphasis on the applicability of the hierarchy model and the influence of advertising on the direction and rate of movement through the different stages of the hierarchy. A distinctive feature of this test is the use of a model, involving time lags, which overcomes many of the causal inference problems that plague field studies.

Specifying the model. The dynamic recursive model displayed below presents a model that is comprised of a system of three interrelated equations. The dependent variables are the three hierarchy variables of interest: brand awareness (cognition), brand attitude (attitude), and brand market share (behavior). The independent variables in each equation are possible causal factors. One such causal factor is contemporaneous advertising exposure. The expectation is that advertising will be significant in the awareness equation but not so in the market share equation. This expectation is not included as a prespecified constraint, but rather exists as a hypothesis to be empirically supported or questioned. The remaining causal factors in each equation are the three hierarchy variables lagged one time period.[2] Because the system of equations does not involve a nonlagged dependent variable among the independent variables the model is termed recursive.[3]

---

[1] Requests for reprints should be sent to David A. Aaker, Associate Professor, School of Business Administration, 350 Barrows Hall, University of California, Berkeley, California 94720.

[2] An index of prices was available for three of the brands. It was treated as a causal variable but did not contribute significantly to the analysis. To economize the presentation the price variable is not reported in this study.

[3] Technically, a recursive model of system is one in which the coefficient matrix of the nonlagged endogenous variables is triangular. Such a requirement is fulfilled, of course, when there are no nonlagged endogenous variables on the right side of the equations. Recursive models have an important statistical property that nonrecursive or interdependent models lack. The parameters can be estimated by applying ordinary least squares to each equation and the resulting parameter estimates are unbiased.

A Dynamic Recursive Model

$$\begin{aligned}\mathrm{AW}(t) &= a_0 + a_1\mathrm{AW}(t-1) \\ &\quad + a_2\mathrm{ATT}(t-1) + a_3\mathrm{MS}(t-1, t) \\ &\quad + a_4\mathrm{ADV}(t) + u_a(t) \\ \mathrm{ATT}(t) &= b_0 + b_1\mathrm{AW}(t-1) \\ &\quad + b_2\mathrm{ATT}(t-1) + b_3\mathrm{MS}(t-1, t) \\ &\quad + b_4\mathrm{ADV}(t) + u_b(t) \\ \mathrm{MS}(t, t+1) &= c_0 + c_1\mathrm{AW}(t) \\ &\quad + c_2\mathrm{ATT}(t) + c_3\mathrm{MS}(t-1, t) \\ &\quad + c_4\mathrm{ADV}(t) + u_c(t)\end{aligned}$$

where:

$\mathrm{ADV}(t)$ = Advertising exposure at time t
$\mathrm{AW}(t)$ = Brand awareness at time t
$\mathrm{ATT}(t)$ = Attitude toward brand at time t
$\mathrm{MS}(t, t+1)$ = Market share of brand during the time period t to $t+1$
$u(t)$ = Error or residual term
a, b, and c terms = Regression coefficients.

Wold has persuasively argued (Strotz & Wold, 1971; Wold, 1954; Wold & Jureen, 1953) that the causal interpretation of a model without such time lags can be ambiguous. The introduction of time lags, on the other hand, makes the model dynamic and tends to expose the direction of causal influence. In this context, if attitude change causes behavioral change, then the dominant causal link is from attitude to behavior, and the coefficient of lagged attitude in the market share equation should be significantly more positive than the coefficient of lagged market share in the attitude equation. If the time lags were not introduced, a causal link might be inferred, but the direction of causation would be in doubt. It should be noted that path analysis (Werts & Linn, 1970; Wright, 1954), another approach to the study of causal networks, usually excludes time delays.

The concept of causal links operating with a delay has been used by Campbell in developing the cross-lag correlation technique (Campbell & Stanley, 1963; Rozelle & Campbell, 1969). The dynamic model used in this study can be viewed as an extension of cross-lag correlation in two respects. First, it expands the causal network from two to three variables. Second, it allows the introduction of control variables (the advertising variable in this case).

The model is also in a distributed-lag format, since in each equation the dependent variable, lagged one period, is included among the independent variables. The coefficient of the lagged dependent variable becomes a measure of the rate at which the influence of the independent variables is dissipated. A large coefficient indicates a slow decay.

The hypotheses suggested in this article were developed primarily from a consideration of the behavior of individuals. Clearly, people differ, and the same ones do not behave identically in all situations. Since aggregate data are used in this study, we are asking which hypotheses have enough general application to emerge in the face of the aggregation. It should be emphasized that conflicting hypotheses can indeed be simultaneously operating in the market.

METHOD

This study focuses on the choice of a brand of instant coffee, a low-involvement decision process.[4] The most attractive feature of low-involvement decisions is that many of the known reasons for behavior appearing to be inconsistent with a previously stated attitude (Wicker, 1971) are minimized. Specifically, there is not likely to be a significant influence from (*a*) personal factors, such as other attitudes or an inability to make or understand the appropriate behavioral response, or (*b*) situational factors stemming from differences in the settings in which attitudes are elicited and choices are made.

The equations in Figure 1 were estimated from data on *aggregate* measures of brand awareness, attitude, and advertising exposure taken from telephone surveys spaced at two-month intervals, and with market share data derived from actual instant coffee sales during the period between the surveys. Completed survey data for 19 time points, covering the period February 1963 to May 1967, were available.

Each of the 19 independent cross-section telephone surveys was based on a national probability sample of 1,200 households who had used instant coffee within the preceding 60 days. Each survey was timed to coincide with the beginning of the two-month audit period. Before an interview began it was necessary to establish that the household was a user of instant coffee and that the respondent made the brand choice decision. Since only 40% of the population satisfied the usage requirement, it was necessary to contact 3,000 households for each survey. Thus the data reported in this study represents the outcome of 57,000 telephone interviews. The basic questions used in this study were unchanged during the 19 surveys.

---

[4] Of course, the level of involvement with instant coffee varies across people. It would be of interest to obtain such a measure on an individual basis and conduct separate analyses for (relatively) high-involvement groups versus (relatively) low-involvement groups. What is meant is that the average involvement is low relative to issues like jobs and to purchases of durables.

The awareness level of each brand at time t, AW(t), was the proportion of households who were able to associate the brand name with the product class without aid. The specific question was: "What are all the different brands of instant coffee (both decaffinated and regular type instant) that you know of that can be bought around here?"

The aggregate *attitude* score for each brand at the t^{th} survey, ATT(t), was based on a weighted average of the proportion of the sample saying the brand was the best at time t and the proportion saying it was "one of several better brands."[5] Since this attitude score is based on comparative rather than absolute judgments of the brands, it can be readily related to market share which describes comparative sales performance.

The advertising exposure variable at time t, ADV(t), was obtained from a three-stage interview-coding procedure. First, each respondent was asked: "What are all the different brands of instant coffee you can recall having seen or heard advertised during the past thirty days, either on radio or TV, in magazines or newspapers?" For each brand mentioned, the respondent was then asked: "What can you recall about the advertising for that particular brand—specifically what are they saying or showing about that brand in their advertising these days?" These recollections were then compared with the actual copy claims being made by the brands at the time of the survey. The comparison was made by trained coders. The advertising exposure variable for each brand, ADV(t), was based on the proportion of respondents who were able to correctly recall one or more specific copy claims being made by each brand.

The aggregate behavior measure was based on the brand and product sales data obtained from bimonthly audits of a national sample of more than 2,000 retail stores. These audits were conducted by a well-known research firm as part of a syndicated service.

During the three-year period covered by this study, the total dollar and unit sales of instant coffee were relatively stable; thus changes in the sales of one brand came primarily at the expense of the other brands in the market. For this reason the aggregate behavior measure is the share of the total sales (or choices) of each brand. This measure, MS($t,t+1$), is the average dollar market share during the period t to $t+1$. Since the market was dominated by five nationally distributed and advertised brands (none with less than a 5% average market share), this meant there were five parallel opportunities to study the relationship of changes in the attitudes of each brand relative to the other brands, with changes in market share. These changes were substantial; the share of the largest brand in the market varied between 32% and 38.5% during the course of the study.

RESULTS

Since scatter diagrams of a sample of the data did not suggest any nonlinearities, no transformations were performed. The model was estimated using ordinary least squares for each of the five brands individually and for all five brands combined. The results are shown in Table 1. When the brands were aggregated, a dummy variable was inserted as the intercept to compensate for the different variable levels at which the brands operated. An assumption of the model when applied to this aggregate datum was that each coefficient was the same for all brands. A gross comparison of the individual brand coefficients and their standard deviations indicated that this assumption was not extremely unreasonable.[6]

The R^2 values for the individual brands generally indicated strong relationships. The lowest R^2 was .30, and in two thirds of the equations it was over .50. Further, in over half of the equations the lagged dependent variable was not significant so the remaining variables were explaining most of the variance. The R^2 for the equations with aggregated data were deceptively high (over .97). The five brands had relatively small variation around different levels reflecting the fact that they had very different average market shares. Thus, the total variance increased enor-

---

[5] The attitude score $\text{Att}(t)_m$, for each Brand m, is based on the following weighted sum of responses:

$$\text{ATT}(t)_m = \text{Proportion of total sample saying Brand m is the "best" in the market}/2.00 + \text{Proportion of total sample saying Brand m is "one of several better brands"}/3.86.$$

The weights are based on the judgments of an independent sample of 40 housewives of the degree of favorableness or unfavorableness of the item on all 11-point scales, using Thurstone's scaling procedure (Edwards, 1957). Theoretically this score can range from 0, if no users of the product thought the brand had any merit, to 50.0 if everyone agreed the brand was the best. Respondents could also choose other statements, including "this brand is acceptable, but not one of the best," or "unacceptable."

[6] The model was also run with the data divided by the appropriate mean for each brand. The results were not affected significantly. Thus, the results using the original data are reported since their interpretation is more meaningful.

TABLE 1

REGRESSION EQUATIONS: ESTIMATING THE DYNAMIC RECURSIVE MODEL

Brand	Dependent variable	Independent variables				C[a]	R^2[b]	Standard error	Durbin-Watson	Sample size
		AW($t-1$)	ATT($t-1$)	MS($t-1, t$)	ADV(t)					
1	AW(t)	.39[c] (1.80)	3.30 (1.20)	−2.01 (1.40)	2.85 (2.42)	21.8 (3.65)	.50	3.96	1.34	19
2	AW(t)	.01 (.06)	−.18 (.11)	2.13 (1.44)	2.6 (3.82)	11.2 (1.6)	.62	1.42	1.52	19
3	AW(t)	.30 (1.06)	.75 (.47)	−.33 (.26)	.16 (.81)	23.2 (2.41)	.42	1.81	1.98	19
4	AW(t)	.26 (.96)	−4.13 (1.30)	.95 (.57)	1.2 (2.6)	33.61 (1.97)	.38	4.67	2.10	19
5	AW(t)	.61 (2.56)	−.08 (1.52)	.08 (.24)	.42 (2.67)	24.5 (1.5)	.57	1.42	1.52	19
All	AW(t)	.37 (3.60)	−.09 (.86)	.33 (.71)	.46 (3.01)	[d]	.977	3.20	1.90	90
1	ATT(t)	−.01 (.68)	.45 (1.84)	.07 (.57)	.13 (.92)	.88 (1.70)	.30	.35	1.84	19
2	ATT(t)	−.10 (.15)	−.18 (.53)	.68 (2.07)	.10 (.68)	.48 (.32)	.30	.39	1.77	19
3	ATT(t)	.11 (1.89)	−.08 (.26)	.09 (.34)	(.03 .85)	−.21 (.11)	.55	.37	1.55	19
4	ATT(t)	.01 (.67)	.16 (.70)	−.03 (.27)	.11 (3.37)	2.51 (2.02)	.54	.34	1.98	19
5	ATT(t)	.23 (1.28)	.02 (.49)	.13 (.51)	.27 (2.3)	−13.5 (1.09)	.51	1.07	2.70	19
All	ATT(t)	.01 (.36)	.03 (1.70)	.26 (3.00)	.10 (3.60)	[d]	.986	.60	2.20	90
1	MS($t, t+1$)	.05 (1.23)	.70 (1.40)	.03 (.18)	.01 (.04)	2.45 (1.80)	.40	.65	1.03	19
2	MS($t, t+1$)	−.05 (1.26)	.35 (2.20)	.76 (5.20)	−.76 (.07)	1.27 (1.86)	.79	.20	1.98	19
3	MS($t, t+1$)	.06 (1.12)	.03 (.14)	.70 (3.30)	.02 (.68)	−.31 (.16)	.80	.32	1.68	19
4	MS($t, t+1$)	.05 (1.27)	.71 (1.27)	.23 (1.05)	−.08 (.80)	−4.34 (1.38)	.34	.72	2.36	19
5	MS($t, t+1$)	.23 (1.77)	.66 (3.35)	.34 (2.07)	−.12 (1.37)	−4.33 (.46)	.78	.83	2.40	19
All	MS($t, t+1$)	.05 (6.05)	.40 (5.77)	.51 (6.92)	.06 (1.63)	[d]	.998	.52	2.34	90

Note. t values are indicated by parentheses. AW(t) = brand awareness at time t; ATT(t) = attitude toward brand at time t; MS($t, t+1$) = market share of brand during time period t to $t+1$; ADV(t) = advertising exposure at time t; AW($t-1$) = brand awareness at time period $t-1$; ATT($t-1$) = attitude toward brand at time period $t-1$; MS($t-1, t$) = market share of brand during the time period $t-1$ to t.

[a] Constant or intercept term.

[b] R^2 are unadjusted.

[c] Regression coefficients.

[d] Dummy variables for each brand were inserted in the aggregate equations.

mously when the brands were combined. The standard error of regression is the more appropriate indicator of the performance with the aggregated set.

A common problem in time-series studies is serial correlation which leads to inefficient coefficient estimates and understated estimates of coefficient variances. The Durbin-Watson statistics shown in Table 1 indicate that this statistical problem is not troublesome in this research. Over half of the values are above the upper limit of the Durbin-Watson test, and none are close to the lower limit at which the hypothesis of serial independence is rejected.[7]

---

[7] This test is biased when lagged dependent variables are used. This bias, of course, is small when the lagged dependent variables are insignificant as they often were in this research, especially in the awareness and attitude equations. Further, several of the equations were run with the dependent variable omitted. The results suggest that if the bias were removed from the test, the hypothesis of serial independence would still not be rejected at the .05 level.

Testing specific hypotheses. The influence of awareness on attitude seems extremely low. In the aggregate, the t value is only .36. Only Brand 3 has a positive coefficient significant at the .10 level, and Brands 1 and 2 have negative coefficients. In contrast, awareness seems to be much more significant in the market share equation. In the aggregate, the t value is 6.05, and all the coefficients except that of Brand 2 are positive with t values exceeding 1.0. This result casts some doubt on the hierarchy concept. However, it is undoubtedly also reflecting the fact that there is a longer time lag between $\text{AW}(t-1)$ and $\text{ATT}(t)$ (a full two-month time period) than between $\text{AW}(t)$ and $\text{MS}(t,t+1)$. It may be that a shorter time lag between $\text{AW}(t-1)$ and $\text{ATT}(t)$ would show a great relationship. Nevertheless, the minimal impact of lagged awareness on attitude and the strong link between awareness and market share is surprising.

As expected, neither the attitude nor the market share variables seemed to influence the awareness variable to any significant degree. The relevant t values in the aggregate case were both below 1.0. It should be noted that Brand 2 has a relatively high-positive-market share coefficient in the awareness equation. Thus, there is some suggestion in the Brand 2 case, especially noting the negative awareness coefficient in the market share equation, that usage is preceding and perhaps causing changes in the awareness level. There are at least two possible explanations for this finding. First, impulse buying may be particularly important for this brand. Second, it might indicate that for this brand the link between awareness and market share influences purchase behavior very quickly. If awareness and market share were affected in the same time period, the awareness would appear to follow the market share change since it was measured at the end of the time period. Brand 2 differed from the others in that it had a higher emphasis on quality. It was also available as a regular coffee. Further, it was the smallest brand studied and did grow slightly during the time periods involved.

The attitude and market share terms that are at the heart of the two major competing hypotheses are considered next. Does an attitude change precede and cause a behavioral change, or does the reverse occur? It seems clear that attitude is a significant influence on market share in this study. In the aggregate, the t value of the attitude coefficient in the market share equation was 5.77, and the t values for Brands 2 and 5 were over 2.2. Only for Brand 3 was the t value low.

In the attitude equation, the conclusion is not so clear. The market share coefficients and the t values are lower. Only Brand 2 had a significantly positive coefficient. These results thus offer some support to the hierarchy-of-effects hypothesis. However, there is certainly evidence of some influence moving in the opposite direction. In the aggregate, the market share t value in the attitude equation was 3.0, and for Brand 2 it was 2.07. Also, the effect of market share on attitude seems higher than that of awareness. However, there is a methodological bias that tends to emphasize the market share to attitude flow of influence. It is caused by the fact that a market share and attitude change in the same time period appear as a market share to attitude flow of influence. With this in mind, the attitude to market share flow of influence predicted by the hierarchy model does seem to appear more dominant in this study.

The advertising variable was expected to influence awareness, and to a lesser extent attitude, but to have little direct impact on market share; these hypotheses seem to be supported. The impact of advertising on market share seems small. It is negative, in fact, for three of the five brands. In contrast, advertising is a strong influence on awareness and attitude. In the awareness equation, there is only one case where the advertising variable is not significantly positive. In the attitude equation, the results for the individual brands are less significant, but the overall results are still strong. Advertising for Brands 1 and 2 seemed to have more impact on awareness than on attitude. It has already been noted that Brand 2 was the smallest studied. Brand 1 was similar in market size to Brand 2, and Brands 4 and 5 were the largest. In this study, advertising seemed to have an equal impact on awareness and attitude. Again it should be noted that these results could partially be reflecting the nature of the time lags actually operating.

The coefficients of the lagged endogenous variables in the recursive equations are also of interest. In general, lagged effects were more important in the market share equations. Thus, the effect of the independent variables on awareness and attitude seems to dissipate rather rapidly, whereas their impact on behavior is of a longer-run nature. This conclusion, where confirmed, suggests that advertising campaigns addressed solely to awareness and attitude may not have the long-term impact they desire unless the campaign's impact is felt fairly soon in market behavior.

DISCUSSION

The empirical results in this study suggest that advertising can directly influence behavior (market share) by working through awareness. There are undoubtedly those to whom brand awareness is an important determinant of purchase for this product class. However, advertising can also affect behavior through its influence on attitude—probably by reaching a different type of buyer. These observations held across the five brands used in this study. Thus, it appeared that all brands had advertising campaigns that affected both awareness and attitude simultaneously.

One surprising result was that the influence of advertising seems to go from awareness directly to behavior and does not operate through the attitude construct. In addition, relatively strong support for the hypothesis that attitude change precedes behavior change was found in this study. Not surprisingly, the reverse effect also emerged, although it was relatively less significant in the aggregate.

It should be noted that this study was conducted under the handicap of significant multicollinearity. Thus, the regression coefficients tend to have higher variances and covariances than are desirable. Fortunately, the *t* test is still appropriate, and conclusions based on it are still valid. However, one must realize that misspecification could distort these findings. Additionally, the data involved only coffee users. In other contexts, particularly when nonusers of a product class are involved, the results may differ.

Two further qualifications concern the generally crude nature of the attitude and advertising effects variables that tend to mask subtle changes in the market and the mature but highly competitive character of the instant coffee market. Other measures and markets may yield different communication processes when subjected to this type of time-series analysis. Nonetheless, the consistency of the effects across five brands enhances the validity of the conclusions.

REFERENCES

CAMPBELL, D. T., & STANLEY, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963.

EDWARDS, A. L. *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts, 1957.

KIESLER, C. A., COLLINS, B. E., & MILLER, N. *Attitude change: A critical analysis of theoretical approaches*. New York: Wiley, 1969.

KRUGMAN, H. E. The impact of television advertising: Learning without involvement. *Public Opinion Quarterly*, Fall 1965, 29, 349–356.

LAVIDGE, R. J., & STEINER, G. A. A model for predictive measurements of advertising effectiveness. *Journal of Marketing*, 1961, 25, 59–62.

MCGUIRE, W. Personality and attitude change: An information-processing theory. In A. G. Greenwald, T. C. Brock, & T. M. Ostrom (Eds.), *Psychological foundations of attitudes*. New York: Academic Press, 1968.

ROZELLE, R. M., & CAMPBELL, D. T. More plausible rival hypotheses in the cross-lagged panel correlation technique. *Psychological Bulletin*, 1969, 71, 74–80.

STROTZ, R. H., & WOLD, H. O. A. Recursive versus nonrecursive systems: An attempt at synthesis. In H. M. Blalock (Ed.), *Causal models in the social sciences*. Chicago: Aldine-Atherton, 1971.

WERTS, C. E., & LINN, R. L. Path analysis: Psychological examples. *Psychological Bulletin*, 1970, **74**, 193–212.

WICKER, A. W. An examination of the "Other Variables" explanation of attitude-behavior inconsistency. *Journal of Personality and Social Psychology*, 1971, **19**, 18–30.

WOLD, H. O. A. Causality and econometrics. *Econometrica*, 1954, **22**, 114 ff.

WOLD, H. O. A., & JUREEN, L. *Demand analysis*. New York: Wiley, 1953.

WRIGHT, S. The interpretation of multivariate systems. In O. Kempthorne et al. (Eds.), *Statistics and mathematics in biology*. Ames: Iowa State College Press, 1954.

(Received March 26, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. [illegible], 287–291

COMPONENTS OF PERCEIVED RISK IN PRODUCT PURCHASE:

A CROSS-VALIDATION

LEON B. KAPLAN
Advertising Research Department,
E. I. DuPont, Wilmington, Delaware

GEORGE J. SZYBILLO AND JACOB JACOBY[1]
Department of Psychological Sciences,
Purdue University

Previous investigations of perceived risk have used different components to estimate the consequence of overall perceived risk. Jacoby and Kaplan (1972) incorporated all these components in the same study and determined which of the components best predicted overall perceived risk for 12 product categories. The present study cross-validated the earlier results. One hundred and four subjects evaluated the physical, psychological, social, financial, performance, and overall perceived risk associated with the purchase of 12 products. Jacoby and Kaplan's (1972) prediction equations were applied to the data to obtain predicted overall perceived risk values for all products. Results of the cross-validation reflected very high agreement with the earlier data.

Since Bauer's seminal statement (1960), many researchers have used the construct of perceived risk to investigate various aspects of consumer behavior. Unfortunately, this construct has not been given a standard definition. For example, in summarizing the early work on perceived risk, Cunningham (1967) defined this concept in terms of two basic components, uncertainty and consequences, with the latter subdivided into two categories, performance and psychosocial consequences. Since then, other investigators (e.g., Perry & Hamm, 1969; Roselius, 1971) have distinguished psychological from social consequences and identified yet other types of consequences (physical safety and financial). As a result, it has been difficult to compare the results across studies of perceived risk.

Accordingly, an investigation was conducted in 1970 (Jacoby & Kaplan, 1972) to determine the interrelationships among the five delineated types of consequences and to determine their individual and collective relationship to overall perceived risk.[2] In general, the results indicated that the five types of consequences explained an average of 74% of the variance in the overall perceived risk measures taken across 12 different products. Performance risk tended to explain more variance than did any other type of consequence.

As has often been noted and recently restated by Heeler and Ray (1972), such findings, though interesting, are unsupported conjectures until and unless they are cross-validated. Unfortunately, most consumer behavior literature has not been cross-validated. Accordingly, this study reports the results of a second study, undertaken two years later and employing different subjects, which cross-validated the original findings.

METHOD

Subjects

The subjects, 104 upperclassmen at Purdue University enrolled in consumer psychology during the spring 1972 semester, were not naive; they had just

---

[1] Requests for reprints should be sent to J. Jacoby, Department of Psychological Sciences, Purdue University, West Lafayette, Indiana 47907.

[2] Detailed operational definitions of the varieties of perceived risk may be obtained from J. Jacoby.

[2] Concurrent with Jacoby and Kaplan's (1972) work which conceptualized risk in terms of five subcomponents, Roselius (1971) identified a sixth variety of risk: "Time loss: when some products fail, we waste time, convenience, and effort getting it adjusted, repaired, or replaced [p. 58]."

TABLE 1

RANKING OF PRODUCTS BY MEAN RISK VALUES FOR EACH TYPE OF PERCEIVED RISK

Product	Overall		Performance		Financial		Social		Psychological		Physical	
	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*
1. Sports car	7.18	1.64	7.26	1.75	7.76	1.43	7.42	2.04	7.32	1.70	5.74	2.32
2. Life insurance	6.50	2.10	6.58	1.99	6.99	1.99	4.04	2.31	4.73	[illegible]	3.67	2.60
3. Color TV	6.17	1.99	6.35	2.09	6.04	1.99	5.82	2.33	4.98	2.30	3.01	1.99
4. Suits	5.91	1.65	5.58	1.93	6.38	1.84	7.13	1.88	6.81	1.89	1.93	1.43
5. Winter coats	5.18	1.93	5.00	2.09	5.54	2.00	6.37	1.88	6.16	2.04	2.46	1.58
6. Dress shoes	4.98	1.84	4.59	1.93	5.04	1.96	5.97	2.16	5.47	2.03	3.02	2.00
7. Deodorants	3.66	2.01	4.25	2.34	3.80	2.07	3.39	2.34	3.81	1.84	3.14	1.99
8. Toothpaste	2.94	1.54	2.81	1.74	2.61	1.47	2.46	1.66	2.23	1.12	3.57	1.96
9. Razor blades	2.92	1.54	3.41	1.91	3.30	2.60	2.04	1.43	2.06	1.31	4.11	2.03
10. Vitamins	2.76	1.61	2.66	1.82	2.88	1.76	1.76	1.10	1.77	1.09	3.63	2.03
11. Aspirin	2.33	1.47	2.37	1.87	1.94	1.07	1.49	.99	1.42	.90	3.54	2.02
12. Playing cards	1.60	1.07	1.65	1.10	1.66	1.02	2.20	1.77	1.75	[illegible]	1.24	.67

completed a two-hour class on the topic of perceived risk, with each risk variety defined. However, they received no information about the earlier study, or the manner in which the five types of perceived risk were related to each other and to overall perceived risk.

Instrument

A questionnaire was developed to measure the amount of each specific type of perceived risk as well as the overall perceived risk these subjects associated with 12 different consumer products. Products were selected to cover a substantial portion of the overall perceived risk continuum. An assortment of health, recreational, and hygienic products, varying along an expensive-inexpensive dimension and including products bought for oneself and for others, highly visible and low-profile items, intimate and nonintimate products, were used. The only restriction was that the product had to be appropriate for both sexes. The number of products was limited to 12 to reduce respondent fatigue and to minimize error.

Each page of the questionnaire examined one aspect of perceived risk (i.e., either overall perceived risk, or one of the five consequences) for one of the 12 products. The top of each page contained the definition of each consequence. The interested reader is referred to Jacoby and Kaplan (1972) for a description of the considerations underlying the development of these specific operational definitions and an elaboration of the procedures employed to identify the five types of consequences.

Analyses

Means, standard deviations, and the intercorrelation matrix were determined for these variables; also, stepwise multiple regression was used to compute the multiple correlations between the components of risk and overall perceived risk for all 12 product categories. As suggested by Mosier (1951) and Lord and Novick (1968), the regression coefficients from the previous study (Jacoby & Kaplan, 1972) were applied to the data of this study to obtain predicted overall perceived risk values for all product categories. These predicted overall perceived risk values were then correlated with the actual overall perceived risk values to obtain cross-validated composite correlations for each product category.

RESULTS

The means, standard deviations, ranges, and rank orders of the risk ratings by individual product category were highly similar to those obtained in the first study.[3] The overall perceived rank order tends to correspond with what would be expected if the products were ordered simply by price alone. In addition, a certain degree of construct validity was established in that the order of risk components was similar within meaningful clusters of products. For example, the rank order of the risk components for the three apparel products (i.e., suit, overcoat, and dress shoes) was the same and went from social risk through psychological, financial, and performance risk to physical risk. Additionally, the rank order of risk components for two of the drug products (toothpaste and aspirins) was the same and went from physical risk through performance, financial, and social risk to psychological risk. These results agree with those obtained from the previous study (i.e., cross-validity). Similar products also show similar risk component hierarchies (i.e., construct validity).

---

[3] Tables with data described but not presented here because of space limitations are available upon request from J. Jacoby, as is the earlier Jacoby and Kaplan study.

Table 1 shows the products ranked according to their mean values on each of the risk components. As in the earlier study, the financial risk column indicates that the mean values diminish gradually until the value for dress shoes; this is followed by a relatively substantial drop in mean value (5.04 to 3.80) for deodorants, after which the means again decrease modestly. While a monotonic relationship between actual price and perceived financial risk could be expected, the relationship appears to be curvilinear. At the upper end of the price continuum (foreign sports car and life insurance) actual dollar value appears less important than relative cost. The most meaningful break (for the sample of college students used) should occur between dress shoes ($15–$35) and (deodorants, razor blades, and other $2–$3 items.), and that seems to be what happened.

An analogous situation would seem to exist for social risk. The largest break in mean values occurs between color TV and life insurance (5.82 to 4.04). A color TV and the products rated above it on social risk (sports cars, suits, coats, and shoes) are all highly visible, as opposed to life insurance and the lower-rated products. These results are similar to those reported earlier.

TABLE 2

INTERCORRELATIONS MATRIX FOR THE FIVE TYPES OF RISK AND OVERALL PERCEIVED RISK CALCULATED ACROSS THE 12 PRODUCT CATEGORIES

Risk	1	2	3	4	5	6
1. Financial	1.00	.67	.76	.23	.59	.76
2. Psychological		1.00	.61	.16	.79	.69
3. Performance			1.00	.33	.58	.79
4. Physical				1.00	.11	.33
5. Social					1.00	.67
6. Overall						1.00

TABLE 3

SUMMARY TABLE FOR PREDICTING OVERALL PERCEIVED RISK FROM ITS COMPONENT VALUES ACROSS ALL 12 PRODUCT CATEGORIES

Step	Variable	*R*	Increase in r^2	Coefficient
1	Performance	.79	.62	.34
2	Psychological	.83	.07	.12
3	Financial	.84	.02	.25
4	Social	.85	.01	.16
5	Physical	.86	.01	.12
	Constant			.18

In the earlier study, performance risk correlated highest with overall perceived risk for 8 of 12 products and second highest for another 3 products. The data from the present study indicate that performance risk correlates highest with overall perceived risk for 10 of the 12 product categories. Both studies thus offer strong support for the practice used by many risk researchers, of employing performance risk as an approximation of overall perceived risk.

Table 2 gives the intercorrelation matrix of risk components and overall perceived risk across the 12 products. As in the previous study, performance risk correlates highest with overall perceived risk (.79) and is followed closely by financial risk (.76). The highest correlation between risk components was that between psychological and social risk (.79), followed by that between financial and performance risk (.76).

Multiple correlations for overall perceived risk ranged from a low of .53 for playing cards to a high of .82 for deodorants. All were significant at $p < .01$. Therefore, 28% to 66% of the variance associated with the overall perceived risk ratings can be accounted for by the five putative varieties of perceived risk employed in this study and first described in Jacoby and Kaplan (1972). Table 3 summarizes the results of trying to predict overall perceived risk from its components across all 12 product categories. The five components accounted for 73% of the variance in overall perceived risk ($p < .01$).

Table 4 presents what is probably the strongest evidence for a substantial degree

TABLE 4

MULTIPLE CORRELATIONS USING FIVE COMPONENT VALUES TO PREDICT OVERALL PERCEIVED RISK FOR FIRST AND SECOND STUDIES WITH CROSS-VALIDATED COMPOSITE CORRELATIONS

Item	Column 1	Column 2	Column 3
	R First study	*R* Second study	Cross-validated composite correlation
1. Aspirin	.75	.75	.72
2. Suit	.71	.57	.53
3. Razor blades	.83	.66	.63
4. Deodorant	.83	.82	.79
5. Foreign sports car	.64	.54	.53
6. 23-inch table color TV	.65	.73	.71
7. Pair of dress shoes	.71	.79	.77
8. Winter coat	.75	.77	.76
9. Toothpaste	.77	.73	.70
10. Life insurance	.65	.76	.74
11. Vitamins	.73	.72	.70
12. Playing cards	.74	.53	.52
13. Overall	.78	.86	.84

of cross-validation. The regression weights from the equations of the previous study were applied to the data of this study and predicted criterion scores for overall perceived risk were obtained. These scores were correlated with the actual observed overall perceived risk criterion scores of the present study and yielded the cross-validated composite correlations which appear in column 3 of Table 4. Column 2 presents the multiple correlations actually obtained in the present study.

The shrinkage in the multiple correlations from the first study (column 1) to the cross-validated composite correlations (column 3) is relatively negligible for aspirin, deodorant, toothpaste and vitamins, and moderate for suits, razor blades, foreign sports car, and playing cards. However, the shrinkage seems minimal when one compares the cross-validated composite correlations for those products to their multiple correlations derived from the optimal weights of this study (column 2). There was actually an increase in the cross-validated composite correlations in comparison to the multiple correlations obtained in the previous study for color TV, dress shoes, winter coats, and toothpaste. However, in no case does the cross-validated composite correlation (column 3) exceed the corresponding multiple correlation obtained in the present study (column 2). They are necessarily lower than the multiple correlations of the present study, for these multiple correlations were obtained with the optimal weights derived from this study. Finally, note that the difference between the predicted (column 2) and obtained (column 3) r_ms never reached or exceeded .04. In most instances these values hover near .02 and suggest a remarkable degree of cross-validation.

DISCUSSION AND CONCLUSION

These results may be considered on two levels: substantive and methodological. Substantively, there are at least three implications. First, the data again indicate that overall perceived risk can be fairly well predicted with the five types of consequences identified earlier (cf. Jacoby & Kaplan, 1972). Second, as found in the earlier study, the data clearly indicate that performance consequences are most predictive of overall perceived risk for most kinds of products. Third, it reconfirms that similar types of products possess similar risk-consequences hierarchies.

Methodologically, the results of the present attempt at cross-validation agree strongly with those of the earlier investigation. Additionally, beta weights based on data from the first study were highly accurate predictors when applied to the data obtained in the present investigation. Thus, the attempt at cross-validation appears successful.

The implications of the present investigation are straightforward. The data from both investigations should (*a*) facilitate comparison of results across studies employing different types of consequences as surrogates for overall perceived risk, and (*b*) provide a basis for understanding and perhaps resolving some of the conflicting findings of other studies.

REFERENCES

BAUER, R. A. Consumer behavior as risk taking. In R. S. Hancock (Ed.), *Dynamic marketing for a* *changing world.* Chicago: American Marketing Association, 1[illegible].

CUNNINGHAM, S. M. The major dimensions of perceived risk. In D. F. Cox (Ed.), *Risk taking and information handling in consumer behavior.* Cambridge, Mass.: Harvard University Press, 1967.

HEELER, R. M., & RAY, M. L. Measure validation in marketing. *Journal of Marketing Research,* 1972, **9,** 36[illegible]70.

JACOBY, J., & KAPLAN, L. The components of perceived risk. In M. Venkatesan (Ed.), *Proceedings, Third Annual Conference, Association for Consumer Research.* University of Chicago, 1972, 382–393.

LORD, F. M., & NOVICK, M. R. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

MOSIER, C. I. Problems and designs of cross validation. *Educational and Psychological Measurement,* 1951, **11**(1), 5–11.

PERRY, M., & HAMM, B. C. Canonical analyses of relations between sociometric risk and personal influences in purchase decisions. *Journal of Marketing Research,* 1969, **6,** 351–354.

ROSELIUS, T. Consumer rankings of risk reduction methods. *Journal of Marketing,* 1971, **35,** 56–61.

(Received May 10, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 292–296

EFFECT OF ANONYMITY ON RETURN RATE AND RESPONSE BIAS IN A MAIL SURVEY

CAROL FULLER[1]

Naval Personnel Research and Development Laboratory, Washington, D.C.

One group of officers and enlisted men to whom the Navy Personnel Survey 71-1 was mailed were instructed to record their service numbers on their answer sheets, while another group were instructed not to identify their answer sheets. Answer sheets were returned more frequently by officers asked to identify their answer sheets. Differences in answers to attitude questions were low in relative frequency and small in magnitude. Pro-Navy statements were endorsed by a higher proportion of the officers who identified their answer sheets; negative statements were endorsed by a higher proportion of the officers who did not identify their answer sheets. None of the anonymous–identified group differences for the enlisted men's questionnaire were large enough to be of practical importance.

Among the potential sources of error in the use of self-report data is the assumed tendency for respondents to distort their answers in a direction favorable to the sponsor of the survey. It frequently has been assumed that an important factor in determining the magnitude of this response bias is whether the respondents are asked to indicate their identity on their answer sheets. However, there are important advantages to be gained from the use of identified questionnaires. When responses can be identified, it is possible to relate those responses to other data concerning the respondents. It is also possible, when respondents are identified, to follow up nonrespondents and thereby avoid bias due to self-selection within a survey sample.

As described by Guilford (1954) and Anastasi (1968), the tendency for respondents to "fake good" on self-report inventories is motivated by the desire to make a favorable impression. This falsification requires both that the respondent be motivated to create a favorable impression and that he know what responses are regarded as favorable (Scott, 1968).

While it generally has been assumed that respondents are less likely to fake good when responding anonymously, previous research using personality inventories and attitude questionnaires has revealed small or no differences in the response given by subjects under anonymous and under identified conditions. This finding is consistent for research in which the same individuals have been tested under both anonymous and identified conditions (Corey, 1937; Fischer, 1946; Gerberich & Mason, 1948; Olson, 1936) and when equivalent groups have been tested (Ash & Abramson, 1952; Hamel & Reif, 1952; Kosen, Kitchen, Kochen, & Stodolosky, 1970; Rosen, 1960).

The direction of the differences that have been reported between anonymous and identified conditions has been consistent with a faking good or social desirability interpretation. Thus, Fischer (1946) found no difference in the mean number of personal problems checked but did find a greater number of "serious problems" checked when the reports were unsigned. Becker and Bakal (1970) found that anonymous–identified conditions had less effect on the responses of subjects who were low in defensiveness than subjects who were high in defensiveness. Rosen (1960) found that an identified group expressed more positive attitudes toward a particular course in which they were enrolled, while there was no difference between the anonymous and the identified groups in their attitudes toward reading. Elinson and Haines (1950) reported finding a higher proportion of pro-military attitudes among an identified group of Air Force and Army enlisted men than among an anonymous group. In a survey of attitudes toward candidates for public office, Benson (1941) found a greater number of "undecided" answers in response to interviews than to secret ballots.

---

[1] Requests for reprints should be sent to Carol Fuller, 3300 South 5th Street, Arlington, Virginia 22204.

This study was conducted to assess the effect of anonymity and identification on the responses of naval officers and enlisted men to a mail survey. Specifically, the two groups of officers and the two groups of enlisted men were compared on (*a*) return rates, (*b*) demographic characteristics, (*c*) the number of blanks on the answer sheets, (*d*) the number of neutral responses, and (*e*) the number of negative and positive attitude statements endorsed by the respondents.

METHOD

The questionnaire used in conducting the study was the Navy Personnel Survey 71-1 (Braunstein, 1972b, 1972c). The Navy Personnel Surveys consist of structured questionnaires administered periodically to collect attitude and opinion data from Navy-wide samples.

Two equivalent groups of officers and two equivalent groups of enlisted men were randomly selected. One group of 6,500 officers and one group of 16,250 enlisted men were asked to identify their answer sheets. These individuals are referred to as identified officer and identified enlisted, respectively. One group of 6,500 officers and one group of 16,250 enlisted men were asked not to identify their answer sheets in any way. These individuals are referred to as anonymous officer and anonymous enlisted, respectively. The post office returned 3,559 questionnaires as undeliverable, leaving 6,173 identified officers, 6,203 anonymous officers, 14,780 identified enlisted men, and 14,785 anonymous enlisted men.

RESULTS

Return Rates

The return rates for each of the four groups are presented in Table 1. Those officers who identified their answer sheets were more likely to respond than were those who did not identify their answer sheets. Among the enlisted men, the return rates were more similar for the two groups.

TABLE 1

RETURN RATES FOR IDENTIFIED–ANONYMOUS OFFICERS AND IDENTIFIED–ANONYMOUS ENLISTED GROUPS

Group	Questionnaires		
	No. administered	No. returned	Proportion returned
Officer			
Identified	6,173	3,147	.51
Anonymous	6,203	2,469	.40[a]
Enlisted			
Identified	14,780	5,293	.36
Anonymous	14,785	5,735	.39[b]

[a] Difference in proportion returned for identified–anonymous officers = .11, $z = 12.3$, $df = 1$, $p < .01$.

[b] Difference in proportion returned for identified–anonymous enlisted = .03, $z = 6.25$, $df = 1$, $p < .01$.

Questionnaire Responses

The proportion of identified respondents endorsing each demographic item was compared to the proportion of anonymous respondents endorsing the item. The (95%) confidence interval for sampling error was subtracted from the difference. Differences less than 5% were considered to be too small to be of practical significance. The only differences (minus sampling error) greater than 5% found between the two officer groups indicated more officers on shore duty (identified officers, 76%; anonymous officers, 59%) and more nonpilot line officers (identified officers, 27%; anonymous officers, 12%) among the identified officer group. The only difference (minus sampling error) greater than 5% between the two enlisted groups indicated that more men in the identified group were not rated (identified enlisted men, 28%; anonymous enlisted men, 19%).

The mean number and the proportion of blank and neutral responses are presented in Table 2. While enlisted men tended to give more blank and neutral responses than did the officers, there were no differences between the identified and the anonymous officer groups or between the identified and the anonymous enlisted groups.

For each alternative answer to each attitude/opinion item, the proportion of identified individuals who chose that alternative was subtracted from the proportion of anonymous individuals who chose that alternative. The (95%) confidence interval for sampling error was subtracted from this difference.

Officers. There were 599 alternative answers for the questions in the officers' questionnaire. The anonymous–identified difference

TABLE 2

MEAN NUMBER OF BLANKS AND NEUTRAL RESPONSES GIVEN BY IDENTIFIED–ANONYMOUS OFFICERS AND IDENTIFIED–ANONYMOUS ENLISTED GROUPS

Group	Blanks		Neutral responses	
	f	$\bar{X}$	*f*	$\bar{X}$
Officer	6,088	2	21,797	7
Identified				
Anonymous	5,187	2	15,729	6
Enlisted				
Identified	36,835	7	43,689	8
Anonymous	36,483	6	45,088	8

exceeded sampling error for 118 alternatives (mean difference = 3%; range = 1%–14%). For 62 of the 118 questions, the anonymous–identified difference did not exceed sampling error for any alternative.

Among the 599 alternatives there were 165 that expressed negative or positive attitudes/opinions (e.g., "Disagree: There are no enlisted poor performers in my immediate work group."). Of these 165 attitude items, 85 were positive statements and 80 were negative statements.

The anonymous–identified difference exceeded sampling error for 20 of the 85 positive statements (1%–10%). For only 2 of these 20 statements did the anonymous–identified difference (minus sampling error) exceed 5% (7% and 10%). Each of the 20 positive statements was endorsed more frequently by the identified group.

The anonymous–identified difference exceeded sampling error for 25 of the statements expressing negative attitudes. For 7 of these 25 negative statements, the anonymous–identified difference (minus sampling error) was greater than 5% (5% or 6%). Each of the 25 negative statements was endorsed more frequently by the anonymous group.

Enlisted. There were 635 alternative answers for the questions in the enlisted men's questionnaire. The identified–anonymous difference exceeded sampling error for 51 alternatives (mean difference = 2%; range = 1%–7%). For 88 of the 120 questions, the identified–anonymous difference did not exceed sampling error for any alternative.

Among the 635 alternatives were 198 statements that expressed negative or positive attitudes/opinions. There were 99 positive and 99 negative statements.

The identified–anonymous group difference exceeded sampling error for 6 of the positive statements (1% or 2%) and for 6 of the negative statements (all = 1%). Each statement was endorsed more frequently by the anonymous group for all but 1 of these 12 statements.

DISCUSSION

The difference in the return rates for the two officer groups could be due either to a heightening of the motivation for response among the identified group or a dampening of response motivation among the anonymous group. Comparison of these return rates with those for the Navy Personnel Survey 70-1 (Braunstein, 1972a; Braunstein & Muldrow, 1972) supports the latter interpretation. All persons were asked to identify their answer sheets for Navy Personnel Survey 70-1. Of the officers who were mailed questionnaires in 1970, 78% responded; 52% of the enlisted men responded in 1970. While the return rates for both of the officer groups and both of the enlisted men's groups in 1971 were lower, the rates for the two enlisted men's groups were depressed about equally. The return rate for the anonymous officer group was reduced even more than the return rate for the identified officer group. Two interpretations of the lower return rate for the anonymous officer group are possible. Certain persons may feel that their opinions are important to the study only when they are identified as their individual opinions. An unintended consequence of anonymity may be a reduction of this perceived value of the specific individual's responses. An alternative interpretation is that the officers tend to feel some pressure to respond and the anonymity instructions reduced this perceived pressure.

The anonymous and the identified officer groups were found to differ on two demographic characteristics, and the anonymous and the identified enlisted men's groups differed on one such characteristic. It is possible that there was an interaction between these characteristics and the identification instructions such that persons with these characteristics were more likely to respond when instructed to identify their answer sheets than when their answers were anonymous. It is also possible that, by chance, more of these particular officers and enlisted men were mailed the identification instructions. There is no evidence available that provides a basis for choosing between these alternative interpretations.

While all of the differences for attitude statements between the anonymous and the identified officer groups were not large (1%–10%), the direction in every case was consistent with a social desirability bias. There were very few differences for attitude statements between the anonymous and the identified enlisted men's groups, and the differences were very small (1%–2%).

Looking at both the lower response rate for the anonymous officer group and the more favorable attitudinal responses of the identified officer group suggests an interpretation of the results in terms of an interaction between the instructions and the presurvey attitudes of the officer groups. This interpretation assumes that individuals with positive attitudes toward the Navy are more likely to return their answer sheets when they are asked to identify their answers than when they are instructed not to identify their answers. This would lead to a higher proportion of individuals in the identified officer group who have positive attitudes toward the Navy and a higher proportion of individuals in the anonymous officer group who have negative attitudes toward the Navy. Since the attitudes of the respondents were not measured prior to the administration of the survey, it is not possible to determine whether the more favorable responses of the identified officer group are due to the influence of the identification instructions or to a greater probability of response among individuals with positive attitudes when they received the identification instructions.

The low frequency and small magnitude of the anonymous–identified differences are consistent with previous research. It is possible that there was little social desirability bias in the responses of either group; that is, both the anonymous and the identified groups responded candidly. Since the content of the survey was not highly personal or sensitive, it is possible that individuals in neither group felt any need to dissimulate. However, previous research indicates little or no difference in the responses of identified and anonymous groups even for highly sensitive information (Ash & Abramson, 1952; Corey, 1937; Fischer, 1946; Kosen et al., 1970; Olson, 1936). King (1970) found no difference between anonymous and identified groups in the proportion of individuals who admitted the use of illegal drugs.

A comparison of the responses given by an anonymous and an identified group, however, does not necessarily indicate the amount of social desirability bias in the responses of the identified group. It seems unlikely that social desirability bias is the result of a deliberate, highly conscious attempt to mislead the researcher. Rather, it has been viewed as a nondeliberate tendency of which the respondent is not aware (Anastasi, 1968; Edwards, 1957). Kosen et al. (1970) found no difference in the response to neutral questions or emotionally threatening questions between a group which rated "possible questionnaire items" and a group which responded orally to the questions in the presence of the experimenter. Items relating to sensitive information may be more likely to elicit social desirability tendencies but these tendencies are apparently evoked equally for anonymous and for identified respondents. If one were interested in determining the amount of social desirability bias in particular questionnaire responses, it would be necessary to employ additional, indirect methods of measurement that are less susceptible to this potential source of error (Campbell & Fiske, 1959; Cook & Selltiz, 1964).

Guarantees of anonymity frequently may be ineffective. Becker and Bakal (1970) suggested that for certain subjects, anonymity instructions may make subjects *more* concerned about the favorability of their responses. Also, it is possible that those subjects who are more predisposed to social desirability bias may be more prone to be suspicious regarding assurances of anonymity. If subjects in the anonymous group are not confident that their answers are anonymous, then there is no real difference in the conditions for response between an anonymous and an identified group.

REFERENCES

Anastasi, A. *Psychological testing.* (3rd ed.) New York: Macmillan, 1968.

Ash, P., & Abramson, E. The effect of anonymity on attitude questionnaire response. *Journal of Abnormal and Social Psychology,* 1952, **47**, 722–723.

Becker, G., & Bakal, D. A. Subject anonymity and motivational distortion in self-report data. *Journal of Clinical Psychology,* 1970, **26**, 207–209.

Benson, L. E. Studies in secret-ballot technique. *Public Opinion Quarterly,* 1941, **5**, 79–82.

Braunstein, C. *Report of enlisted findings: Navy Personnel Survey (NPS) 70-1.* (WSR 73-1) Washington, D.C.: Naval Personnel Research and Development Laboratory, 1972. (a)

Braunstein, C. *Report of enlisted findings: Navy Personnel Survey (NPS) 71-1.* (WSR 72-7) Washington, D.C.: Naval Personnel Research and Development Laboratory, 1972. (b)

Braunstein, C. *Report of officer findings: Navy Personnel Survey (NPS) 71-1.* (WSR 72-9) Washington, D.C.: Naval Personnel Research and Development Laboratory, 1972. (c)

Braunstein, C., & Muldrow, T. W. *Report of officer findings: Navy Personnel Survey (NPS) 70-1.* (WSR 72-8) Washington, D.C.: Naval Personnel Research and Development Laboratory, 1972.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 1959, **56**, 81–105.

Cook, S. W., & Selltiz, C. A multiple-indicator approach to attitude measurement. *Psychological Bulletin,* 1964, **62**, 36–55.

Corey, S. M. Signed versus unsigned attitude questionnaires. *Journal of Educational Psychology,* 1937, **28**, 144–148.

Edwards, A. L. *The social desirability variable in personality assessment and research.* New York: Dryden, 1957.

Elinson, J., & Haines, V. T. Role of anonymity in attitude surveys. *American Psychologist,* 1950, **5**, 315.

Fischer, R. P. Signed versus unsigned personal questionnaires. *Journal of Applied Psychology,* 1946, **30**, 220–225.

Gerberich, J. B., & Mason, J. M. Signed versus unsigned questionnaires. *Journal of Educational Research,* 1948, **42**, 122–126.

Guilford, J. P. *Psychometric methods.* (2nd ed.) New York: McGraw-Hill, 1954.

Hamel, L., & Reif, H. G. Should attitude questionnaires be signed? *Personnel Psychology,* 1952, **5**, 87–92.

King, F. W. Anonymous versus identifiable questionnaires in drug usage surveys. *American Psychologist,* 1970, **25**, 982–985.

Kosen, D., Kitchen, C., Kochen, M., & Stodolosky, D. Psychological testing by computer: Effect on response bias. *Educational and Psychological Measurement,* 1970, **30**, 803–810.

Olson, W. C. The waiver of signature in personal reports. *Journal of Applied Psychology,* 1936, **20**, 442–450.

Rosen, N. A. Anonymity and attitude measurement. *Public Opinion Quarterly,* 1960, **24**, 675–679.

Scott, W. A. Attitude measurement. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology.* (2nd ed.) Vol. 2. *Research methods.* Reading, Mass.: Addison-Wesley, 1968.

(Received February 21, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 297–301

CONTRIBUTION OF PERSONALIZATION TO MAIL QUESTIONNAIRE RESPONSE AS AN ELEMENT OF A PREVIOUSLY TESTED METHOD[1]

DON A. DILLMAN[2] AND JAMES H. FREY

Washington State University

Two experiments reported the effects of varying degrees of personalization on response rates to mail questionnaires. In Experiment 1, half of an alumni sample received three personalized cover letters with a university questionnaire, while the other half received three unpersonalized letters. The return rate was higher in the personalized condition, although the later introduction of a highly personalized method in both personal and not personal conditions did not increase response. In Experiment 2, nearly a methodological replication, half of a sample of the general public received a different questionnaire without prior contact, while the other half received the questionnaire after contact by telephone. The call did not increase response rates but did stimulate earlier return of questionnaires.

A recently reported method (Dillman, 1972), found successful in stimulating high response to mail questionnaires, depended on systematic manipulations of virtually all aspects of the procedure such as questionnaire format, nature of appeals, and follow-up mailings. The use of this method twice resulted in a high-response rate (75%) to a 12-page (109–150 item) questionnaire from the general public, traditionally a population from which it is hard to elicit high response. One aspect of this concerted application of techniques was personalized communications. However, the extent to which it alone contributed to the high response was not determined. One purpose in the present study is to present results from two experiments which tested the effect of personalization when all of the other factors of the mail technique noted above were controlled.

Personalization is defined as the process of creating a belief on the part of the respondent that he is receiving the researcher's individual attention. More specifically, personalization refers to how one is communicated with and not just whether the communication occurs. Ways frequently used to accomplish personalization include individually typed letters, handwritten letters or postscripts, individual saluations, hand-applied signatures, telephone calls to each respondent, and some types of follow-up.[3] The experiments reported in this study deal with the following types of personalization: (*a*) individually typed names, inside addresses, and personal salutations in conjunction with individually signed letters (versus preprinted salutations and signatures), and (*b*) prior contact (versus no prior contact) by telephone.

A number of researchers have argued that personalization is one of the important elements of a method that increased response (Reeder, 1960; Roeher, 1963; Tallent & Reiss, 1959). However, experimental data were not available to support their contentions. Experiments by others have resulted in increases of five to eight percentage points as a result of personalization (Linsky, 1965; Longworth, 1953; Simon, 1967). Two other experiments have shown no significant positive effects from personalization (Clausen & Ford, 1947; Kawash & Aleamoni, 1971). Finally, there have been some instances in which personalization actually produced a lower response than did an impersonal procedure (Andreasen, 1970; Simon, 1967).

---

[1] This article is a revision of a paper read at the Pacific Sociological Association in Phoenix, Arizona, May 3–5, 1973. Funds for this research were provided by the Social Research Center and Project 0031 of the Agricultural Research Center, Washington State University, Scientific Paper #3942.

[2] The authors wish to express their sincere appreciation to James F. Short, Jr., and Linda Smith for the considerable assistance they gave to this endeavor.

Requests for reprints should be sent to Don A. Dillman, Chairman, Department of Rural Sociology, Washington State University, Pullman, Washington 99163.

[3] Follow-up letters may or may not be a form of personalization. If a follow-up is sent only to those not responding to an earlier mailing, and that fact is indicated, for example, "we have not yet received your questionnaire," some degree of personalization is achieved.

These mixed results from past research, plus the fact that the studies differed in many ways (e.g., topic and length), make it difficult to determine what the effect would be if a researcher added a personalization procedure to his study. A second problem associated specifically with the studies in which personalization had a positive effect was that response rates ranged from only 20% to 60%. Unless response rates of a magnitude as great or greater than that achieved in the Dillman (1972) research can be consistently obtained, the mail questionnaire technique will be of limited value to social scientists. If personalization is not found to push relatively high response even higher, then it can be discarded in favor of mass mailing procedures with considerable savings of time and money.

Our tests of personalization procedures are concerned with the effect of varying degrees of personalization rather than with just the effects of its presence and its absence. The first experiment tests the effects of mail personalization techniques, and the second experiment tests the effects of using personalization by telephone in addition (rather than as an alternative) to personalization by mail.

EXPERIMENT 1—PERSONAL SALUTATION AND REAL SIGNATURE IN COVER LETTERS

Method

The first experiment was conducted in conjunction with a sample survey of 891 Washington State University alumni done during the summer of 1972. The procedural details were identical with those reported by Dillman (1972) except as necessitated by the different topic. The questionnaire (8 pages, 78 items) sought responses of the alumni to questions dealing with their feelings and concerns about the policies of Washington State University.

The personalization procedure was implemented in the following fashion: (*a*) To achieve the appearance of being individually typed, cover letters were typed on letterhead stationery and reproduced by a multilith process which resulted in a letter image very comparable to the one which would have been originally typed. (*b*) Names, street addresses, and dear M__. ______ were individually placed on each letter using the same typewriter. (*c*) The project director's signature was individually signed in blue ink (for contrast) with a ballpoint pen to leave an impression on the paper as evidence of a "real" signature. One half of the alumni received cover letters produced in this manner. The control group received entirely multilithed letters with the salutation of "Dear Alum" and with a black preprinted signature. The experiment was continued with a follow-up postcard (an applied signature versus a preprinted signature), and a second follow-up letter produced in the same manner as the original cover letter. Thus, one group received three personalized communications, the other group none.

A subexperiment was included in the fourth mailing which consisted of a replacement questionnaire and a cover letter sent by certified mail. Half of the nonrespondents from each of the original groups received a letter done in identical fashion to the previous one they had received. The remainder of each group received a "highly personalized" letter that was individually typed. Thus, the two original personalization treatments were expanded into four: (*a*) personal throughout all four mailings, (*b*) personal in three mailings and highly personalized in the fourth mailing; (*c*) impersonal throughout all four mailings; and (*d*) impersonal in three mailings and highly personal in the fourth mailing. A discerning respondent could distinguish between the "personal" and "highly personal" in part because of the indentations produced by the typewriter. The individually typed letters were produced on a magnetic tape selectric typewriter (MT/ST). After being placed on tape, the letters could be efficiently produced at the rate of 20 per hour by the operator who added inside addresses and salutations to each.

Results

The personalization treatment used in the first three mailings did make a difference in the results. The proportion of questionnaires returned after these mailings was nearly 9-percentage points greater for the personalized group ($\chi^2 = 8.29$; $p \leq .01$). On the basis of these findings we conclude that the personalization procedures are an important part of the method and their omission would adversely affect response.

The results of the subexperiment included in the fourth mailing were mixed. The addition of the highly personalized method did not increase response. The response rates were virtually the same for both the group receiving the nonpersonal treatment through all four mailings and the nonpersonal subgroup receiving the switch to the highly personalized treatment in the fourth mailing (45.9% versus 45.3%). The personal subgroup given the high-personalization treatment in the fourth mailing actually had a lower response than did the rest of its group which received the personalized treatment through all four mailings (50% versus 73.3%).

These findings support the notion that one cannot recover from a low-response rate with a last-ditch effort at personalization. One possible explanation for the findings which show no difference in response rates is that by the time a person had received his fourth mailing he took very little note of the specific communication, having guessed what it was about. Although the letters were different in terms of personalization procedures, the questionnaire and return envelopes were not. His decision to respond or not to respond at that time may have been based on the carry-over impact of personalization or the nonpersonalization procedures through previous mailings and not the fourth mailing per se. If so, then we have seen even a stronger argument for early implementation of personalization.

Another possible but unexplored effect of personalization might be to increase the speed with which questionnaires are returned. This is desirable for two reasons. First, if returns come in faster, follow-up is less expensive because fewer people need to be contacted. Second, reducing the retrieval time lessens the chance of an uncontrolled event affecting the response quality. Surveys need to be implemented as rapidly as possible on issues for which attitudes are changing (e.g., candidate preferences in an election) or issues which are subject to the intrusion of some major event (e.g., the assassination of a public figure during a study of opinions on gun control). Thus, increasing the speed of returns should, in the long run, improve data quality. A plot of the cumulative day-by-day returns for each of the experimental groups revealed that the respondents who received the personalized communication did return their questionnaires more quickly, but only to a slight extent. A difference of two to three percent was consistent through the fourth mailing. This was hardly enough to justify the personalization procedure for that reason alone.

TABLE 1

RESPONSE RATES FOR LETTER PERSONALIZATION EXPERIMENTS APPLIED TO ALUMNI STUDY, OVERALL AND BY TREATMENT

Treatment	No. of returns		%
	No. mailed	No. completed	
Response to first three mailings			
Personal	458[a]	353	77.1
Not personal	469	322	68.7
Unidentified	—	3[b]	—
Totals	927	678	73.1
Response to fourth mailing			
Personal	45[a]	33	73.3
Highly personal	46	23	50.0
Subtotals	91	56	61.5
Not personal	61	28	45.9
Highly personal	64	29	45.3
Subtotals	125	57	45.6
Totals	216	113	52.3
Totals			
Personal	444[a]	409	92.1
Not personal	447	379	84.8
Unidentified	—	3[b]	—
Grand total	891	791[b]	88.8

[a] This represents the number of questionnaires which could have been completed. It is somewhat lower than the number of questionnaires mailed because of incorrect addresses, deaths and because of the inclusion of persons on the original mailing lists who were "supporters" of the university, but not alumni. It was further adjusted downward, as reflected on the totals column; due to the receipt of information on additional ineligible respondents *after* the fourth mailing had been sent.

[b] This figure includes respondents whose personalization treatment could not be identified because they had torn the identification number from their questionnaire.

EXPERIMENT 2—A PRIOR TELEPHONE CALL

Method

The telephone experiment builds on the mail experiment previously reported in Experiment 1. A full set of mail personalization procedures, from the first mailing to the last, was applied to all the respondents. The telephone was used, experimentally, to increase the dose of personalization received by part of the respondents. The object was to determine the effects of the increased increment of personalization.

A sample ($n = 696$) of the general public of the state of Washington was systematically drawn from all telephone directories in the state. The questionnaire was 12 pages long and required 89 responses. It was believed that the topic, similar to that of the alumni study, had little salience to most of these persons, and it would be difficult to secure a high response to the lengthy and detailed questionnaire.

Telephone contact was designed to encourage mail response. The procedure followed in each contact was to (*a*) explain the nature of the study; (*b*) create interest by asking a few questions on the topic, selected because of their presumed interest to the respondent; (*c*) explain that we wanted to send them a mailed questionnaire; (*d*) emphasize that they, individually, were of great importance to the study; and (*e*) answer their questions. Clearly a telephone conversation can convey the fact of individual attention to the respondent.

The experimental procedure was as follows: The sample was divided into two halves—one group to receive a mail questionnaire without prior contact, the other to be telephoned before receiving the questionnaire. Only two thirds of the group to be telephoned was actually interviewed. Twenty-eight could not be reached even after 8 to 10 attempts in addition to checks with the operator to determine if the numbers were working and correct. Of those contacted, nearly 23% refused to be interviewed.

Results

Response rates for each group were as follows: (*a*) those interviewed by telephone, 75.2%; (*b*) those who refused over the telephone, 50%; (*c*) those who could not be reached by telephone, 45.2%; and (*d*) the control group, 64.5%. People interviewed on the telephone were significantly more likely than were the control group to return completed questionnaires ($\chi^2 = 7.43$; $p \leq .05$). However, the overall response rate for the complete experimental group was virtually the same (66.9%) as that for the control group. Thus, we must raise the question as to whether the telephone made any difference at all. It seems plausible that those people most likely to return a questionnaire are also likely to submit to a telephone interview. We may have talked by phone with a selective subgroup of the sample who would have been just as likely to return a mail questionnaire without the call as with it.

The "high personalization" subexperiment applied to the fourth mailing of the alumni study was also run on the sample of the general public. Half of the nonrespondents in each experimental group were randomly selected to receive the individually typed letters. The remainder received the less personal multilithed letters. Again, the results were not significant. Of those receiving the highly personalized letter, 24.8% returned the completed questionnaires compared to 21.5% who received the less personalized version. These findings reinforced the conclusion of the mail only experiment that a last-ditch high-personalization attempt is unwarranted.

However, the speed of returns was increased rather dramatically. Eighty percent of those interviewed by telephone, who eventually returned the questionnaire, had done so prior to the second follow-up attempt. This compares to only 52% of the control group. The group interviewed by telephone returned 90% of all questionnaires within 25 days of the first mailing. The time elapsed before the control group had passed the 90% mark was nearly double, or 48 days. Even when the no-contact and refusal subgroups who responded more slowly than did the group interviewed were included in the experimental group, the speed of response was still much faster than that for the control group. Seventy-five percent of the total experimental group respondents had returned their questionnaires before the third mailing as compared to 52% of the control group.

DISCUSSION

Personalization effects may have substantial significance for making the mail questionnaire method a viable technique for the collection of social science data. The experiments were with questionnaires of considerable length under conditions of already high response. Other procedures thought to be important on the basis of previous research had already been applied. The results showed that the application of personalization raised high-response rates even higher. An economically produced multilithed letter with the salutation in matching type and a real signature, which together give the "look" of an individually written letter, increased response rates significantly. Although a return rate of nearly 85% may be acceptable, one of just over 92% is even more gratifying, particularly inasmuch as the number of "hard-core" holdouts has been reduced in this particular study by one half.[4] An additional dose of personalization, the telephone call, did not increase the response rate for a separate sample but did significantly increase the speed of reply.

Implications of these experiments extend beyond the personalization procedures used. A further test of the basic method previously reported by Dillman (1972) also has been provided. The general public is among the most difficult of groups from which to get a high response from mail questionnaires. This is because of the difficulty of forming group specific appeals or taking advantage of feelings of reciprocity which have so often played a role in other studies reporting techniques designed to increase responses (Scott, 1961). Another factor for making the general public such an extremely difficult group from which to generate response is that the differences in characteristics such as age, education, and belief structures are at nearly the maximum, creating further difficulties in communication. On these grounds, it is expected that methods of stimulating response should be more effective on specialized homogeneous groups than the general public. The fact that the response rate for the group of alumni receiving the complete treatment (92.1%) was higher than that received by the three samples of the general public (75% in the previously mentioned two response rates reported by Dillman (1972) and 65% in the response rate reported in this study) seems to confirm this observation.[5]

The goal of efforts to improve mail questionnaire response is to increase the viability of the technique for research. It is noteworthy to observe that the response level obtained in the alumni sample approaches what is generally a very satisfactory level for personal interview studies. The geographical dispersion of the Washington State University alumni (43 states and nine foreign countries) would have made personal interviews impossible. Thus, instead of considering the mail questionnaire to be cheaper but an undesirable alternative, it may be considered in some instances as a cheaper but desirable alternative. Its use in such situations can add significantly to our capabilities for surveying difficult to interview populations.

REFERENCES

ANDREASEN, A. R. Personalizing questionnaire correspondence. *Public Opinion Quarterly*, 1970, **34**, 273–277.

CLAUSEN, J. A., & FORD, R. N. Controlling bias in mail questionnaires. *Journal of the American Statistical Association*, 1947, **11**, 497–511.

DILLMAN, D. A. Increasing mail questionnaire response in large samples of the general public. *Public Opinion Quarterly*, 1972, **36**, 254–257.

KAWASH, M. B., & ALEAMONI, L. M. Effect of personal signature on the initial rate of return of a mailed questionnaire. *Journal of Applied Psychology*, 1971, **55**, 589–592.

LINSKY, A. Experimenting in inducing response to mail questionnaires. *Sociology and Social Research*, 1965, **49**, 185–189.

LONGWORTH, D. S. Use of a mail questionnaire. *American Sociological Review*, 1953, **18**, 310–313.

REEDER, L. G. Mailed questionnaires in longitudinal health studies: Problem of maintaining and maximizing response. *Journal of Health and Human Behavior*, 1960, **1**, 123–129.

ROEHER, A. G. Effective techniques in increasing response to mailed questionnaires. *Public Opinion Quarterly*, 1963, **27**, 299–302.

SCOTT, C. Research on mail surveys. *Journal of the Royal Statistical Society*, 1961, **124**, 143–205 (series A).

SIMON, R. Responses to personal and form letters in mail survey. *Journal of Advertising Research*, 1967, **7**, 28–30.

TALLENT, N., & REISS, W. J. A note on an unusually high rate of returns from a mail questionnaire. *Public Opinion Quarterly*, 1959, **23**, 579–581.

(Received April 30, 1973)

---

[4] This should not be construed as meaning personalization should reduce by half of the number of holdouts in any study where personalization is used, rather it underscores the fact that personalization made a considerable dent in a group of potential respondents who were resistant to all of the other techniques applied as a means of increasing response. It is as important to note what happens to the number of holdouts as well as how much the response rate per se is increased.

[5] It must be pointed out, however, that the educational level of the alumni group was obviously quite high, and that this factor has been shown to be associated positively with response (Longworth, 1953). Secondly, as previously noted, the questionnaire sent to the alumni was slightly shorter than those sent to the general public.

Journal of Applied Psychology
1974, Vol. 59, No. 3, 302–306

SCALING PEER RATINGS:

AN EXAMINATION OF THE DIFFERENTIAL VALIDITIES OF POSITIVE AND NEGATIVE NOMINATIONS

GARY G. KAUFMAN[1]
Internal Revenue Service, Washington, D.C.

JAMES C. JOHNSON[2]
Tennessee State Department of Personnel,
Nashville, Tennessee

Peer ratings obtained by the nomination technique are typically scaled by algebraically combining the top nominations with the bottom nominations. It is thus implicitly assumed that top and bottom ratings are equally valid measures of performance. This assumption was tested in two studies of Reserve Officer Training Corps (ROTC) cadets (*N*s = 113, 87) by correlating several kinds of peer rating scales with a number of criteria. The results suggest that negative nominations contribute little to peer rating scales and that the frequency of positive nominations may be the most useful index.

Impressive evidence has accumulated regarding the validity and reliability of peer ratings in many different situations (Hollander, 1965; Kubany, 1957; Roadman, 1964; Tupes, 1957; Weitz, 1958). Two methods of obtaining peer ratings have been widely used. Most common is a nomination procedure in which raters are required to choose, from among their peers, a specified number of persons who are "highest" on some performance continuum and an equal number of persons who are "lowest." Because of its simplicity, this method is especially suited to situations in which a large number of persons are to be rated. The second method is a complete or partial ranking of peers that may be accomplished by several procedures such as simple rank ordering, alternation ranking, or paired comparisons. Ranking methods are typically used when relatively few persons are rated.

In either case, a scale is usually formed by summing algebraically the highest (positive) and the lowest (negative) nominations received by each ratee or by obtaining an average of the combination of positive and negative ratings. These procedures are based on the implicit assumption that combining positive and negative evaluations to form a single scale best reflects the relevant information contained within a set of ratings. The present study tested this hypothesis, for the case of the nomination method, by comparing the correlations between peer ratings scaled by different methods and several other measures of performance.

STUDY 1

Method

Subjects in the initial study were 113 members of the 1969 senior class of the University of Tennessee Army Reserve Officer Training Corps (ROTC). Each cadet was asked to nominate, from a list of names provided him, the five persons in the class who were most effective as officers and the five who were least effective. The nominated names were to be listed in rank order, with the most effective listed first, the next most effective listed second, and so on. The rank ordering was required to provide weights necessary for several of the scaling procedures under investigation. An individual rated most effective was given the score +5, the next most effective, +4, and so on. Among those nominated least effective, the lowest ranking individual was given the score −5, the next least effective, −4, and so on through −1.

The ratings were then scaled by seven different methods: (*a*) the algebraic sum of the weighted nominations; (*b*) the total number of nominations received; (c) the frequency of most effective (positive) nominations; (*d*) the frequency of least effective (negative) nominations; (*e*) the algebraic sum

---

[1] The authors are grateful to the officers and men of the University of Tennessee Army Reserve Training Corps for their generous and willing participation in this study; and we are indebted to E. E. Cureton and Michael Gordon for their comments on an earlier draft of this article.

Requests for reprints should be sent to Gary Kaufman, Room 1513, Internal Revenue Service, 1111 Constitution Avenue, N.W., Washington, D.C. 20224.

[2] Now at the Tennessee State Department of Personnel, 1401 Andrew Jackson State Office Building, Nashville, Tennessee 37219.

TABLE 1

CORRELATIONS BETWEEN PEER RATING SCALES AND ROTC AND SUMMER CAMP CRITERIA

(N = 113)

Criteria	Algebraic sum	No. ratings	f		$\bar{X}$	$\bar{X}$	
			Positive	Negative		Positive	Negative
Distinguished military student	.50[a]	.50	.60	−.17	.49[a]	.40	−.01
End camp leadership	.53[a]	.06	.42	−.44	.48	.35	−.07
Military proficiency	.30	.10	.31	−.25	.20	.10	−.07
Drill performance	.40	.38	.44	−.13	.42	.34	−.17
ROTC class standing	.26	.21	.28	−.14	.28	.34	−.01
Original rank	.62	.59	.71	−.19	.63	.42	−.18
Present rank	.55[a]	.58	.66	−.18	.55[a]	.38	−.13

Note. r = .19 significant at .05 level.

[a] Significantly different from f positive at .01 level (Hotelling, 1940).

of weighted nominations divided by the number of nominations received; (f) the algebraic sum of weighted positive nominations divided by the number of positive nominations received; and (g) the algebraic sum of the weighted negative nominations divided by the number of negative nominations received.

Performance criteria were then collected from the cadets' past performance records in the university ROTC program and at a six-week summer camp. These criterion measures were: (a) distinguished military student, an award made to about 20% of the cadets based on ROTC grades, academic grades, and recommendations from several Army officers; (b) end camp leadership ratings, a graphic rating scale combined with a trait checklist completed by the cadets' summer camp platoon leaders; (c) military proficiency, an objective composite measure based on performance at the rifle range, in compass and map reading courses, physical fitness, and a written examination on military knowledge; (d) drill performance, a rating of the cadets' abilities to conduct military drill and leadership in field situations rated by cadet and regular Army officers; (e) ROTC class standing, a ranking of cadets by ROTC grade point average; (f) original rank, military rank assigned at the beginning of a cadet's senior year; (g) present rank, the cadet's military rank at the end of his senior year.

Results

The correlations between the peer rating scales and the performance criteria are shown in Table 1. These relationships, by visual inspection of the scatter plots, appeared to be reasonably linear and homoscedastic. Intercorrelations among the criteria are shown in Table 2. The frequency of positive nominations emerged as the best predictor of the various criteria, predicting six of the seven criteria at least as well as, and generally better than, did the other scales. The criteria are reasonably independent; initial and final military rank, however, have an extremely high intercorrelation. Taken as a whole, these data suggest that the negative nominations add little relevant variance to peer evaluations.

These results, however, required replication to ascertain that differences between the correlations are reliable, and that the pattern can be generalized to other criterion measures. A second study was therefore undertaken, employing a new sample and different performance criteria. It was also possible, in the second study, to collect peer ratings prior to the subjects' participation in summer camp.

STUDY 2

Method

A nomination procedure identical to that of the first study was used, with the exception of the instructions to rank order the nominees, the results of which contributed little to the validity of the ratings in the first study. The subjects were 87 cadets in their junior year; performance criteria were obtained the following year after their attendance at summer camp. The summer camp criteria had been completely revised, and consisted of three measures: (a) performance, a graphic rating of the cadets' leadership abilities in various summer camp situations by the platoon leaders completed at least

TABLE 2

CRITERION INTERCORRELATIONS ($N = 113$)

Criterion	DMS	End camp leadership	Military proficiency	Drill performance	ROTC class standing	Original rank
End camp leadership	.36	—				
Military proficiency	.31	.26				
Drill performance	.39	.07	.31			
ROTC class standing	.31	.09	.12	.45		
Original rank	.56	.47	.18	.40	.28	
Present rank	.49	.36	.11	.40	.28	.92

Note. $r = .19$ significant at .05 level. DMS = distinguished military student.

four times during the six-week camp; (*b*) personal characteristics, a graphic rating by platoon leaders on 15 descriptive statements, such as "knows how to handle men"; and (*c*) situations, a graphic rating by platoon leader indicating his willingness to allow the cadet to represent him in various types of situations.

The peer rating scales were (*a*) the algebraic sum of the positive and negative nominations; (*b*) the frequency of positive nominations; (*c*) the frequency of negative nominations; and (*d*) the algebraic sum of positive and negative nominations divided by the number of positive and negative nominations (mean rating).

Results

The correlations between the peer rating scales and the three criteria are shown in Table 3. Frequency of positive nominations was again the best predictor of cadet performance. Intercorrelations between criteria were moderate: (*a*) personal characteristics and performance (.51); (*b*) situations and performance (.43); and (*c*) situations and personal characteristics (.40).

In Table 4 are three reliability estimates for top, bottom, and total number of nominations. Values in the first column are intraclass correlations calculated from the weighted nominations contained in the first study ($n = 113$). The calculation yields an estimate of the average reliability of the rating.

The coefficients in the second column of Table 4 were calculated by randomly splitting the raters into two groups, by correlating the frequencies of nominations received for top, bottom, and total across the two halves, and by correcting these correlations by the Spearman-Brown formula. These calculations were based on data from the first study.

The third column of coefficients was calculated using rate-rerate correlations over a six-month period. The data for these correlations came from a third study ($n = 134$), not described in this article (Kaufman, 1972).

The split-half and rate-rerate estimates of reliability were very similar despite the use of different raters and different methods of calculating the coefficients. The intraclass estimate was surprisingly divergent from the other estimates. The intraclass coefficients, however, were calculated on *weighted* nominations, and the split-half coefficients were calculated on the frequencies of nominations: thus, the intraclass coefficients reflected differences between raters in assigning relative positions among the top five and the bottom

TABLE 3

CORRELATIONS BETWEEN PEER RATING SCALES ($N = 89$) AND REVISED SUMMER CAMP CRITERIA

Criteria	Peer rating scales			
	Algebraic sum	f		$\bar{X}$
		Positive	Negative	
Performance	.30	.42	−.03	.25[a]
Personal characteristics	.27	.36	.05	.26
Situations	.37	.43	.16	.35

Note. $r = .21$ significant at .05 level.

[a] Significantly different from f positive at .01 level (Hotelling, 1940).

five, while the split-half reflected rater agreement in assigning individuals to one of two classes, the top five or the bottom five.

DISCUSSION

The results of the two investigations clearly contradict the assumption that an algebraic sum or average of positive and negative nominations yields the most valid prediction index. Of the several procedures investigated in this study, the most useful was simply the frequency of positive nominations. These results are similar to those obtained by Webb (1957), who reported a biserial correlation of .87 between an algebraic sum of the highest three - lowest three peer ratings and a dichotomy of those receiving highest ratings and those receiving no ratings, suggesting that the negative nominations were of little value. Correlations of the peer ratings with other performance measures, however, were not obtained.

As a possible explanation of this phenomenon, Doll and Longo (1962) offered the hypothesis that low ratings are contaminated by reactions of raters to antisocial behaviors which are unrelated to performance criteria. To test this hypothesis, they experimentally removed the low ratings which were given for antisocial reasons and found that the validity of the peer rating was enhanced. Their procedure, however, was fairly complicated and inapplicable for most practical situations.

Our conclusion is also similar to one reached by Berkshire and Highland (1953) who compared several forced-choice rating methods. A forced-choice evaluation method containing only favorable statements correlated more highly with their criteria than did methods using both favorable and unfavorable statements.

More recently, Gordon (1970) has reported evidence of a differential accuracy phenomenon in videotaped interview situations. Interviewers perceived the "desirable" incidents of behavior more accurately than the "undesirable" incidents. Raters may thus be able to identify examples of effective performance more accurately than behaviors contributing to ineffective performance.

In practice, the use of positive nomination methods removes the cause of the frequent complaint that the peer rating asks a rater to "cut his buddy's throat" (Webb, 1955), and may also counter the potential efforts of group members to "save" those men previously given low ratings by giving them high ratings at the next opportunity. Weitz (1958) and Mayfield (1970) report that raters readily accept a positive nomination procedure; interviews carried out in three companies revealed no complaints about the rating system.

The results of this study further support the expectation that peer rating techniques may be a particularly valuable aid in identifying potentially effective managers. The results of Hollander's (1965) study, also carried out in a military setting, suggest that peer ratings may be usefully obtained as early as three weeks after hiring a trainee, and thus provide an inexpensive alternative to more costly methods of assessment.

TABLE 4

RELIABILITIES OF PEER RATING SCALES

Rating	Correlations		
	Intraclass for weighted nominations ($n = 113$)	Corrected split-half ($n = 113$)	Rate-rerate ($n = 134$)
Top	.64	.82	.86
Bottom	−.02	.59	.60
Total	.90	.64	.66

REFERENCES

BERKSHIRE, J. R., & HIGHLAND, R. W. Forced-choice performance ratings—A methodological study. *Personnel Psychology*, 1953, **6**, 355–378.

DOLL, R. E., & LONGO, A. A. Improving the predictive effectiveness of peer ratings. *Personnel Psychology*, 1962, **15**, 215–220.

GORDON, M. E. The effect of the correctness of the behavior observed on the accuracy of ratings. *Organizational Behavior and Human Performance*, 1970, **5**, 366–377.

HOLLANDER, E. P. Validity of peer nominations in predicting a distant performance criterion. *Journal of Applied Psychology*, 1965, **49**, 434–438.

HOTELLING, H. The selection of variates for use in prediction, with some general comments on nuisance parameters. *Annals of Mathematical Statistics*, 1940, **11**, 271–283.

KAUFMAN, G. G. *An examination of the nomination paradigm of peer ratings* (Doctoral dissertation, University of Tennessee) Knoxville, Tenn.: University Microfilms, 1972, No. 72-27, 479.

KUBANY, A. J. Use of sociometric peer nominations in medical education research. *Journal of Applied Psychology*, 1957, **41**, 389–394.

MAYFIELD, E. L. Management selection: Buddy nominations revisited. *Personnel Psychology*, 1970, **23**, 377–391.

ROADMAN, H. E. An industrial use of peer ratings. *Journal of Applied Psychology*, [illegible], **48**, 211–214.

TUPES, E. C. *Relationship between behavior trait ratings by peers and later officer performance of USAF officer candidate school graduates* (Tech. Rep. AFPTRC-TN-57-125) Lackland Air Force Base, Tex.: Air Force Personnel Research Center, October 1957.

WEBB, W. B. The problem of obtaining negative nominations in peer ratings. *Personnel Psychology*, 1955, **8**, 61–63.

WEITZ, J. Selecting supervisors with peer ratings. *Personnel Psychology*, 1958, **9**, 25–[illegible].

(Received February 28, [illegible])

Journal of Applied Psychology
1974, Vol. 59, No. 3, 307-312

AN EMPIRICAL COMPARISON OF THE RELATIVE EFFECTS OF RATER RESPONSE BIASES ON THREE RATING SCALE FORMATS

ROBERT F. BURNASKA[1]
General Electric Company, Portsmouth, Virginia

THOMAS D. HOLLMANN[2]
Wayne State University

Smith-Kendall type behaviorally anchored scales for derived performance dimensions (Format 1), scales for the same dimensions but without the behavioral anchors (Format 2), and scales for dimensions selected on an a priori basis (Format 3) were compared on the basis of susceptibility to rater response biases. Leniency error and composite halo error were present in all ratings; there was no evidence of relative or absolute halo errors in any ratings. There was some evidence that the use of scales for derived dimensions reduced leniency error and increased the amount of variance attributable to ratee differences. The scale reliabilities of the three formats were also determined. A discussion of the feasibility of obtaining relatively independent scales for several job performance dimensions was included.

Numerous attempts have been made to alleviate the response style problems inherent in having one individual rate another by using various formats (e.g., forced-choice, paired comparison) and/or various raters (e.g., peers, self). Although each proposed innovation guards against one or more rater errors, as often as not it will give rise to other serious problems that are unique to the proposed method: Innovations in rating, although plentiful, are likely to result in robbing Peter to pay Paul.

The response style problems, as well as the occasional benefits inherent in rating procedures, can be discussed by breaking down the total rating variance in an analysis of variance framework (Guilford, 1954). In a situation where several raters rate several ratees on several categories, one obtains the following sources of variance:

Raters
Ratees
Raters × Ratees
Categories
Raters × Categories
Ratees × Categories
Raters × Ratees × Categories

The rater main effect is indicative of leniency error (Guilford, 1954). Leniency, a misnomer for this response bias, is analogous to what Cronbach and Gleser (1953) refer to as an elevation when discussing profile comparisons. This error occurs when different raters use the scales differently in an overall way. Some raters consistently give relatively high ratings, while others consistently give relatively low ratings. These differences are, for the most part, consistent across scales as well as across ratees. While the intrarater differences may be valid, they are meaningless without assumptions and corrections for comparisons of persons rated by different raters.

The ratee main effect is composed of true ratee differences across scales and an overall bias which Guilford (1954) calls "absolute halo" and Blumberg, DeSoto, and Kuethe (1966) call "consensual halo." The "absolute" or "consensual" halo bias occurs whenever raters erroneously, and en masse, rate certain ratees higher and others lower. There is no way to statistically separate the valid and biased components. In a situation where the raters are familiar with the ratees' relevant behaviors, however, rater agreement is more likely to represent convergent validity rather than consensual halo error.

A Rater × Ratee interaction is evidence for what Guilford (1954) calls "relative halo" and Blumberg et al. (1966) call "individual halo." This error occurs when some ratees are given high ratings across categories by some raters and low ratings across categories by other raters, but there is no agreement or consensus among raters.

---

[1] Requests for reprints should be sent to Robert F. Burnaska, Corporate Personnel Research, General Electric Company, College Boulevard, Portsmouth, Virginia 23705.

[2] Now at General Electric Company, Fort Wayne, Indiana.

A category main effect represents a composite of true category differences for the ratees and an overall rater bias for rating some categories high and others low. The latter component has the effect of restricting the range of at least some of the scales and thereby reducing the scales' reliabilities and potential utility for correlational purposes.

The Rater × Category interaction describes variance due to raters' idiosyncratic tendencies to use category scales at different levels. This has never been shown to be a serious problem for the utility of rating scales.

The Ratee × Category interaction represents the extent to which the raters, as a group, make ratee-category discriminations. This source of variance is large when each ratee's category profile differs from the other ratees' profiles and when each ratee's category profile is similar across raters. This source of variance should be high since the agreement among raters adds convergent validity to the ratings and the differential profiles of ratees yield maximum information about their relative strengths and weaknesses.

The three-way interaction is, of course, error variance, in the analysis and in concept.

The desirable sources of variance, then, are the ratee main effect and the Ratee × Category interaction. They provide information as to how the ratees differ, both quantitatively and qualitatively, and this is the purpose of all rating scale procedures. Any other sources of variance contribute response style errors of one type or another.

The behaviorally anchored rating scales developed by Smith and Kendall (1963) have been proclaimed as a major stride toward the improvement of rating scales. They were reportedly acceptable to raters, conscientiously used, and it has been felt that the methodology involved in the development of the scales reduces the danger of unwanted error variance attributable to rater response styles.

The reliabilities of the scales have been determined by examining interrater agreement on the assignment of behavioral anchors to scale positions of categories. High-scale reliabilities have been reported by Smith and Kendall (1963) and Folgi, Hulin, and Blood (1971), but only moderate reliabilities have been reported by Zedeck and Baker (1971). Moderate interrater reliabilities have been reported (Hakel, 1971; Landy & Guion, 1970; Maas, 1965) when the Smith-Kendall scales have been used to rate individuals. In addition, several investigators (Dunnette, 1970; Hakel, 1971; Landy & Guion, 1970; Zedeck & Baker, 1971) have found the Smith-Kendall scales to be susceptible to halo error, and Zedeck and Baker (1971) found that they yielded little evidence for construct validity. Maas (1965), however, found that interrater agreement on scaled-expectation rating scales was significantly superior to agreement on adjective rating scales, and Campbell, Dunnette, Arvey, and Hellervik (1973) found that behaviorally anchored scales yield less method variance, less halo error, and less leniency error than do Likert-summated rating scales.

To date, no systematic comparative test has been conducted by evaluating sources of variance to determine if behaviorally anchored scales are both less susceptible to rater response biases and more reliable than more conventional rating scales. This study fills the void by comparing three types of rating scale formats to determine which of the three yields the most reliable ratings with the fewest rater errors. The three formats are (*a*) the Smith-Kendall type, behaviorally anchored scales (Format 1), (*b*) scales for the same dimensions as Format 1 but with adjective rather than behavioral anchors (Format 2), and (*c*) a set of scales measuring performance categories selected on an a priori basis (Format 3).

A very likely reason for the shortage of comparative studies of rating scales is that it is difficult to find situations in which a large number of potential raters have had an adequate opportunity to observe the behavior of a group of ratees, since it is the former which constitutes sample size. This study, by necessity, compromises the desirability of a large number of raters for the suitability of the overall situation and, therefore, some of the results are suggestive rather than conclusive.

METHOD

Constructing the Scales

To construct rating scales for Format 1, a slight modification of the Smith-Kendall procedure was followed.

First, 33 undergraduates at a large Midwestern university were asked to provide critical incidents (Flanagan, 1949) illustrative of very effective and very ineffective classroom-teaching behaviors displayed by college instructors and professors whom they had encountered during their college careers. Each student was asked to provide three positive critical incidents (very effective classroom-teaching behaviors) and three negative ones (very ineffective classroom-teaching behaviors). For each critical incident reported, students were asked to name and define the behavioral category of classroom-teaching behavior of which the critical incident was an example; for example, if the ineffective critical incident pertained to an instructor failing to compensate for his foreign accent, the behavioral category might be "verbal presentation." Behavioral categories of classroom-teaching behaviors were also obtained independently of any critical incidents from another group of 15 undergraduates.

For each critical incident reported, a single summary statement was derived to reflect the essence of the central behavior for that incident. Terminology and meaning as originally reported were retained. These summary statements composed the pool of behavioral statements from which scale anchors were selected.

Next, a group consisting of 17 undergraduates, 10 psychology graduate students, and 5 psychology professors, were presented with the 135 derived critical incident summary statements and the 13 behavioral categories. A brief definition of each behavioral category was provided. This group was asked (*a*) to assign each of the summary statements to the behavioral category which they thought the summary statement was an exemplar, and (*b*) to rate, on a 9-point horizontal scale, the teaching effectiveness of the behaviors represented by each of the critical incident summary statements.

Results were tallied, and for each behavioral statement for which there was high agreement (65% or more of the respondents) as to category placement, the mean and standard deviation of the teaching effectiveness ratings were computed. Those critical incident summary statements within each behavioral category that had differing means and small standard deviations were then chosen for use as anchors for the scale values of Format 1. Choosing those statements with a wide dispersion of means assured coverage of different scale effectiveness levels for each behavioral category; small standard deviations assured agreement on the effectiveness level of each statement chosen. The means and standard deviations of the chosen behavioral anchors for each category were similar, and no problem was encountered in obtaining anchors for the midpoints of the scales.

Each of the critical incident summary statements was then reworded into a behavioral expectation statement. For example, if the critical incident summary statement was "The instructor lectured only from the text," the behavioral expectation statement would be "could be expected to lecture only from the text." The behavioral expectation statements were then applied as anchors to the appropriate scale positions of a 9-point vertical scale for the behavioral category to which they pertained. The behavioral categories used were those that had relatively large numbers of behavioral expectation statements. If few behavioral expectation statements were assigned to a given behavioral category, the category was considered to be either unimportant or nonratable. The resultant behavioral categories for Format 1 were "knowledge of material," "rapport with the students," "verbal presentation," "social consciousness," and "student evaluations."

Format 2 consisted of 9-point vertical scales for the same behavioral categories as used in Format 1, but the brief definitions of the behavioral categories and the scales' behavioral expectation anchors were omitted. These were omitted to determine the extent to which such anchors produce differences in the susceptibility of Formats 1 and 2 to rater response biases. The two extremes and midpoint of these scales were anchored by "very effective," "very ineffective," and "average effectiveness," respectively.

Format 3 consisted of behavioral categories obtained from a teaching effectiveness rating form published by Stoops and Rafferty (1961). The behavioral categories were "knowledge of subject," "understanding pupils," "fairness in marking," "interest in pupils," and "human qualities." The number of behavioral categories was the same as used in Formats 1 and 2, a 9-point vertical scale was again used for each behavioral category, and the extremes and midpoint of the scales were anchored by "very effective," "very ineffective," and "average effectiveness," respectively. This a priori format is representative of many graphic rating scales currently in use, and was used in the present study so that the ratings, rater errors, and reliability of this format could be compared with those produced by the other two systematically derived formats.

Administering the Scales

The raters were first-year graduate students of psychology. The ratees were three associate professors of psychology whom the raters had had in succession during their first year as graduate students. The rating was conducted in a university classroom on the last class day of the third quarter of the regular academic year.

Raters were randomly assigned to use one of the three rating formats. Because of unforeseen circumstances (e.g., illnesses, dropouts, etc.) that prevented equal number of raters for each format, 12 raters used Format 1, nine used Format 2, and nine used Format 3. The order of rating the three professors was counterbalanced for the formats, and all ratings were made independently. Confidentiality of the

TABLE 1

TWO-WAY ANALYSES OF VARIANCE

Source of variance	Format 1		Format 2		Format 3	
	F	Estimates ω^2	F	Estimates ω^2	F	Estimates ω^2
Replications across ratees						
Rater	4.09**	.07	3.54**	.05	7.64**	.19
Category	13.49**	.11	22.79**	.23	14.18**	.19
Rater × Category	.38	.00	.39	.00	.71	.00
Replications across categories						
Rater	1.61	.04	1.26	.02	4.54**	.17
Ratee	26.02**	.27	19.39**	.27	13.29**	.15
Rater × Ratee	1.62	.04	1.51	.04	1.27	.02
Replications across raters						
Ratee	29.92**	.28	19.87**	.27	22.38**	.15
Category	6.10*	.10	8.32**	.21	14.18**	.19
Ratee × Category	1.42	.01	2.04**	.03	.72	.00

* $p < .05$.
** $p < .01$.

ratings and the individual raters was maintained at all times.

For the purpose of determining the effects that the behavioral expectation anchors and brief definitions of the behavioral categories had on the ratings, those who made initial ratings with Format 2 were contacted three weeks later and asked to rate the professors again, using Format 1. This was done to determine whether the presence or absence of the behavioral expectation anchors and brief definitions of the behavioral categories would effect the susceptibility of these two formats to rater response biases.

The reliability of the scales of each format was estimated by following the analysis of variance techniques set forth by Lindquist (1953). By ignoring differences between raters, the test–retest reliability coefficient of the mean rating of each format was obtained.

RESULTS

Following the method of analysis set forth by Guilford (1954), the ratings of the three formats were analyzed by treating them as replications across ratees, categories, and raters, respectively, and thus obtaining three two-way analysis of variance factorial designs for each format in which each source of variance is treated as a random factor. Following this procedure, the three two-way analyses of variance for each format are shown in Table 1.

For Format 1, the significant sources of variance are the category effect and the ratee effect, plus the rater effect when collapsing across ratees. Estimates of omega-squared (Hays, 1963) reveal that the ratee effect is the largest source of variance, accounting for approximately 27% of the total. The fact that the rater effect is significant in one situation and nonsignificant in the other is inconsequential since this effect accounts for a relatively small proportion of variance.

The significant sources of variance for Format 2 are the three main effects; rater, ratee, and category, plus the Ratee × Category interaction. The category and ratee effects are the largest, each accounting for more than 20% of the total variance. The interaction effect is relatively unimportant, accounting for only 3% of the variance.

The three main effects are all significant for Format 3; there are no significant interactions. In each case, these significant sources account for just less than 20% of the total variance.

By treating each main effect as a variance, or a measure of the extent to which the means for the various levels of the factor vary around the grand mean of the factor, one can test the significance of the differences between the main effects of the formats by homogeneity of variance F tests. While none of these tests was significant, yielding no statistical support for discussing differential contributions of any main effect in the three formats, the omega-squared values of these sources of variance suggest that there are important differences between the formats which do not emerge as statistically significant solely because of the relatively small sample sizes. The evidence suggests that using Formats 1 and 2 reduces the variance attributable to the rater effect and increases the variance attributable to the ratee effect, and that using Format 1 also reduces the variance attributable to the category effect.

The reliability of the scales of Format 1 is .78, the reliability of the scales of Format 2 is .81, and the reliability of the scales of Format 3 is .63. Unfortunately, because of the small number of raters for each format, the standard errors of the reliability coefficients are so large that it is relatively impossible to meaningfully demonstrate significant differences between the reliabilities of the scales.

Correlations between similarly named scales for Formats 1 and 2 were obtained to yield additional evidence of the effects that the presence or absence of the behavioral expectation anchors and brief category definitions had on the ratings of these formats. The overall correlation between the two formats is .61 ($p < .01$). The Pearson product-moment correlation coefficients between the category pairs, which are all significantly different from zero at alpha $= .01$; are: Knowledge of Material, +.74; Rapport with the Students, +.62; Verbal Presentation, +.71; Social Consciousness, +.54; and Student Evaluations, +.54.

DISCUSSION

The ratings from all three formats contain leniency error and composite halo error; none contain absolute or relative halo error. While there is some indication that the systematically developed scales reduce the amount of leniency error and increase the amount of variance attributable to ratee differences, there is no evidence for the superiority of any one format.

On the basis of the omega-squared estimates, both Format 1 and Format 2 reduce leniency error, suggesting that the use of the scales which emerge from the methodology employed are more meaningful to raters than are the scales selected on an a priori basis and, therefore, that rater differences are reduced. It is very likely, in fact, that this reduction in the rater component accounts for the increase in the ratee component. As scales become more meaningful to raters and more related to observable behaviors rather than interpretations of behaviors, the behavioral differences between ratees account for more variance in the ratings, and the interpretative differences among the raters account for less.

The one distressing finding is the overwhelming presence of composite halo error. It is inconsequential that the Ratee × Category interaction, the *absence* of which indicates the presence of this error, is significant for Format 2. This effect accounts for only 3% of the variance and can hardly be considered meaningful. The ubiquity of composite halo both in the present study and in other research (e.g., Dunnette, 1970; Landy & Guion, 1970; Zedeck & Baker, 1971) is alarming. The purpose of obtaining performance ratings on several dimensions is to gather information concerning individuals' patterns of strengths and weaknesses. To the extent that the several scale ratings are correlated, another indicant of composite halo, this objective of using the scales is not met.

Maybe it is time to raise the question, How realistic is the objective of obtaining relatively independent scales for several dimensions of job performance? It is impossible to escape the fact that each individual scale of almost any set of rating scales is influenced by some type of pervasive element. Halo error, most would agree, is the result of a global evaluation biasing the ratings on each performance category in the same direction, the result of which is inflated scale intercorrelations. A major proposed benefit of the Smith-

Kendall behaviorally based scales is that they focus the raters' attention on behaviors relevant to each performance dimension and away from a global evaluation. However, by increasing the focus on behavior, the global evaluation may be replaced by something just as global: a pervasive motivational component. If behavior is a function of a multiplicative relationship between undifferentiated drive and specific habit strengths for a given stimulus situation, then drive, or motivation, will have a constant effect on all behaviors at a given time. If work motivation is relatively constant over time, we should expect a correlation between behaviors on a variety of performance dimensions. Perhaps efforts to measure absolute ability levels independently of actual performance, plus a separate assessment of the pervasive motivational component, will prove a more fruitful source of usable information. Perhaps a pervasive motivational dimension, like Spearman's "g," is a fact of life and has to be tolerated.

The test–retest reliability reported in this study for the Smith-Kendall scales is somewhat larger than previously reported (Hakel, 1971; Landy & Guion, 1970; Maas, 1965) interrater reliabilities for these scales. Although the reliabilities of the three formats could not be meaningfully compared for significant differences, perhaps a repetition of the present design with larger sample sizes will provide more conclusive results.

Thus we see again that format modifications and improvements in scale development techniques are not the final remedy to the criterion problem. Other approaches to improve ratings, such as rater training, may prove to be equally or more fruitful alternatives.

REFERENCES

BLUMBERG, H. H., DESOTO, C. B., & KUETHE, J. L. Evaluation of rating scale formats. *Personnel Psychology*, 1966, **19**, 243–259.

CAMPBELL, J. P., DUNNETTE, M. D., ARVEY, R. D., & HELLERVIK, L. V. The development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology*, 1973, **57**, 15–22.

CRONBACH, L. J., & GLESER, G. C. Assessing similarity between profiles. *Psychological Bulletin*, 1953, **50**, 456–473.

DUNNETTE, M. D. A behavioral approach for describing managerial effectiveness. Paper presented at the Bowling Green Symposium, December 1970.

FLANAGAN, J. C. A new approach to evaluating personnel. *Personnel*, 1949, **26**, 35–42.

FOGLI, L., HULIN, C. L., & BLOOD, M. R. Development of first-level behavioral job criteria. *Journal of Applied Psychology*, 1971, **55**, 3–8.

GUILFORD, J. P. *Psychometric methods*. New York: McGraw-Hill, 1954.

HAKEL, M. D. Similarity of post-interview trait rating intercorrelations as a contributor to interrater agreement in a structured employment interview. *Journal of Applied Psychology*, 1971, **55**, 443–448.

HAYS, W. L. *Statistics for psychologists*. New York: Holt, Rinehart & Winston, 1963.

LANDY, F. L., & GUION, R. M. Development of scales for the measurement of work motivation. *Organizational Behavior and Human Performance*, 1970, **5**, 93–103.

LINDQUIST, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953.

MAAS, J. B. Patterned scale expectation interview: Reliability studies on a new technique. *Journal of Applied Psychology*, 1965, **49**, 431–433.

SMITH, P. C., & KENDALL, J. G. Retranslation of expectations: An approach to the construction of unambiguous anchoring of rating scales. *Journal of Applied Psychology*, 1963, **47**, 149–155.

STOOPS, E., & RAFFERTY, M. L. *Practices and trends in school administration*. Boston: Ginn, 1961.

ZEDECK, S., & BAKER, H. T. Evaluation of behavioral expectation scales. Paper presented at the Midwestern Psychological Association, Detroit, May 1971.

(Received May 24, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 313-320

MAGNITUDE ESTIMATIONS OF EXPRESSIONS OF FREQUENCY AND AMOUNT[1]

BERNARD M. BASS,[2] WAYNE F. CASCIO,[3] AND EDWARD J. O'CONNOR[4]

University of Rochester

Magnitude estimation was employed to find the numerical equivalents of 39 expressions of frequency ranging from *never* to *always*, and 44 expressions of amount ranging from *none* to *all*. The results were generalizable across three age occupation levels and unaffected by topic importance. Geometric means and appropriate variance measures are provided for each expression as well as suggestions for anchors for 4- to 9-point scales. The percentage of overlap in judgments for adjacent points on scales are also given. Results are related to earlier work on scaling, and the utility of present approach is indicated.

In applied psychology, we make heavy use of judgments of frequency and judgments of amount. The merit rater judges how much or how often the behavioral statements apply to the employee he is rating; the respondent to the attitude questionnaire judges how much or how frequently the statements reflect his feelings; the personality assessee judges how much or how often the statements are true about himself. Nevertheless a limited number of adverbial and adjectival modifiers are available for expressing the range of extensivity of such judgments in frequency (never to always) and in amount (none to all).

Many other aspects of behavior, sensation, or experience can be described, such as their goodness, strength, and potency (Osgood, Suci, & Tannenbaum, 1957) in terms of frequency and amount. Thus, a scale evaluating "goodness" can be couched in terms of frequencies of occurrence such as "usually good" or "rarely good" or in terms of amounts of occurrence such as "fairly good" or "very good."

Following Fechner's logarithmic law of the relation between stimulus and sensation, Thurstone (1927) introduced subjective scaling procedures for equal interval scaling. Quantitative values were assigned to observed behavior, attitudinal statements, sensations, or experiences in the hope of achieving scales of equal intervals with an arbitrary zero. For purposes of merit rating, for instance, this procedure made possible the identification of the scale weights of 724 statements about the goodness of performance of any job occupant. These median weights of meritorious performance were relatively invariant to the particular sample of judges employed to establish the weights (Uhrbrock, 1950).

At about the same time, Hemphill and Coons (1950) identified 42 adverbs of frequency or extent in connection with the development of the Leader Behavior Description Questionnaire (LBDQ). A small group of staff members completed paired comparisons among these adverbs against the criterion of how much each expressed frequency or extent. The modal rank reached by each adverb was noted. In any pairing with other adverbs, "always," for example, was always judged greater of the pair in how much it expressed frequency or extent. "Never" was always judged of lesser extent in pairings with other adverbs.

Some years earlier, Simpson (1944) had asked 335 high school and college students to indicate the "number of times in a 100" was signified by each of 20 adverbs and adverb phrases. They were asked to give a range of estimates for each answer. Hakel (1958) repeated Simpson's work with 100 Minnesota students. He noted a high degree of agreement between the medians of his samples and those of Simpson as well as large individual differences in response to many of the adverbs. Nevertheless, such individual differences varied considerably from one adverb to the next.

---

[1] This research was supported by the Personnel and Training Research Programs, Psychological Services Division, Office of Naval Research. It was done in connection with the development of scales designed to investigate alternatives that make work more attractive.

[2] Requests for reprints should be sent to Bernard M. Bass, Graduate School of Management, University of Rochester, Rochester, New York 14627.

[3] Now at Florida International University, Miami, Florida.

[4] Now at the University of Akron, Ohio.

Sheppard (1954) determined for 78 British subjects the mean distance on a 7-inch line represented by the phrases and adverbs, very bad, bad, rather bad, and so forth. But much more comprehensively, Cliff (1959) looked at adverbs such as slightly, somewhat, rather, very, etc., as modifiers of a variety of adjectives such as evil, immoral, nice, lovable, etc. Reliability of paired comparison judgments was found to be .999. The multiplying values of "intensity" of each adverb for three college student samples from Wayne, Princeton, and Dartmouth universities were obtained. The high degree of stability across the three samples was striking evidence of the relative invariance of such adverb modifiers.

Cliff's success with a multiplicative model for adverbs of intensity argues strongly for both the theoretical as well as practical utility of a multiplicative rather than an additive approach to adverbs as modifiers of frequency and amount (Hakel, 1958; Hemphill & Coons, 1950; Sheppard, 1954; Simpson, 1944). As a consequence, magnitude estimations using Stevens's (1966) procedures seemed to be a more accurate way of standardizing expressions of frequency and amount, a way which better fitted the true nature of relations between stimulus and subjective experience. Such standardization, if sufficiently invariant for pools of different individuals, could then be used by investigators to build any kind of behavioral rating scale. Objective quantities of frequency and amount could be readily associated with subjective modifiers according to Stevens's (1971) law that equal stimulus ratios produce equal sensations.[5]

Bass (1968) initiated such an effort for a questionnaire survey. The frequency scale he used was based on a magnitude estimation study of 28 adverbs of frequency. Each of 71 undergraduate students had been asked to assign a number of his own choosing to "sometimes," then to indicate what number would best fit each of the 28 other adverbs. On the average, if "sometimes" was set at 1.000, "always" was seen as 2.533 times as frequent; "very often" was 2.093 times as frequent; "fairly often" was 1.683 times as frequent, "seldom" was .425 times as frequent and "never" was .00[illegible]. Thus, these particular 6 adverbs of the 28 studied bore an approximate relation to each other of 5:4:3:2:1:0 and were selected as the response alternatives for a questionnaire.

The purpose of the present investigation was to obtain reliably established geometric means and standard deviations of magnitude estimations of reasonably exhaustive lists of expressions of frequency and amount used to modify attitudinal ratings. Also we set out to see if such results were insensitive to ratings of important, as opposed to unimportant, issues for the raters. It was hypothesized that judges generate one set of ratios of extensivity when rating issues of consequence to them and a different set when judging issues of little importance to them. Only if the hypothesis was rejected would a single standardized list be possible.

METHOD

The method employed was that of ratio scaling. The distinguishing characteristic of a ratio scale is that a true or absolute zero point can be defined; meaningful ratios can therefore be derived. Ratio scaling uses magnitude estimation[6] (Stevens, 1966,

---

[5] The claims for the superior validity and power of magnitude estimation scaling have been questioned due to the fact that if judges are asked to equate differences rather than make ratio judgments, a logarithmic transformation of magnitude estimation scales emerges. Intuitively, we counter that the evidence is less likely to hold for psychological judgments such as word meanings, compared to physical stimuli such as brass weights. Moreover, comparing two differences $(X_1 - X_2)$ and $(Y_1 - Y_2)$ involves judging four things, while magnitude estimation reduces the problem to assigning numbers against a standard anchoring number.

[6] The equal interval status of magnitude scales has been questioned and we have to agree that magnitude estimation and partition scaling give discordant results on prothetic intensity continua. However, on metathetic qualitative dimensions the two approaches give results that are in good agreement. The question then seems to be: "Do our scales of frequency and amount represent prothetic or metathetic continua?" If the latter, no problem exists

1971) in which any value is assigned to a referent concept, and then all other stimuli are judged in relation to the referent concept. For example, if an individual attaches the value 30 to the word "sometimes" he might assign 15 to the word "seldom" if he felt that seldom represented one half as much as sometimes. Likewise he might assign the value 300 to the word "always" if he felt that always expressed 10 times as much as the word sometimes.

Subjects

The subjects were 175 males and females drawn from three populations: night school master of business administration (MBA) students (most of whom were working), adult undergraduate students, and high school juniors.

Design

The subjects scaled 39 expressions of frequency and 44 expressions of amount. Instructions to each subject read as follows:

> People frequently use words such as sometimes, rarely, seldom and never. Assume someone has just said to you, "I *sometimes* think the Viet Nam war should continue until we win."
>
> On the attached form, please assign a number to what you conceive "sometimes" to mean. You may use any number greater than or equal to zero (0). Please place the number on the line next to "sometimes."
>
> Then, using the number you have assigned to "sometimes" as a standard, please assign a number to each of the other words on the form indicating each word's value relative to sometimes. Again, use any number greater than or equal to zero (0). The number placed next to each word or phrase should reflect what you feel that word or phrase means when compared with "sometimes." For example, if you assign a value of 50 to "sometimes," you might assign the value 100 to any other word or phrase which you felt represented twice the frequency of "sometimes" and a value of 5 to any word or phrase that seemed to represent only 1/10 the frequency of "sometimes."
>
> You may use any whole or decimal number greater than or equal to zero (0), just as long as you feel it represents the numerical value of a word or phrase when that word or phrase is compared with "sometimes."

Instructions used the referent "some" instead of "sometimes" when judgments of amount were to be made instead of judgments of frequency. The referent concepts "sometimes" (frequency) and "some" (amount) remained constant throughout the experiment.

The 38 other expressions of frequency and 43 expressions of amount were presented in one of five different orders so as to guard against order effects.

The subjects were randomly assigned to one of two conditions:

Condition 1: expressions of frequency with important topic
expressions of amount with unimportant topic

Condition 2: expressions of amount with important topic
expressions of frequency with unimportant topic.

To ensure agreement with what the experimenters considered topics of importance (air pollution and the Viet Nam war) and unimportance (the amount of rainfall in Nepal, worms in the street after a rain storm), the subjects were asked to rate each topic on a 5-point Likert scale of importance from "extremely important to me" to "extremely unimportant to me." To make more meaningful experimental comparisons, each subject's data were analyzed only if a particular topic was ranked either in the top or bottom third of the distribution of importance-unimportance for that topic. The subjects who ranked topics in the midrange of importance-unimportance were deleted from the analyses.

Results

Results pertinent to the central purpose of the study are presented in Tables 1 and 2. Table 1 presents the means and standard deviations of the 39 expressions of frequency, and Table 2 presents the means and standard deviations of the 44 expressions of amount.

In the first analysis a 2 × 3 repeated measures multiple analysis of variance (MANOVA) was performed using the 39 expressions of frequency as dependent variables.

---

since all agree that for these dimensions the partition scale represents a straight linear segment of the magnitude scale and that the intervals are therefore equal (Stevens, 1971). If, on the other hand, our scales of frequency and amount are prothetic or intensity dimensions the answer is less clear. Stevens (1971) still argues that the magnitude scale has equal intervals while others point out that its lack of linearity with partition scales indicates the inequality of its intervals. We cannot resolve this issue since it's not clear how one draws the fine line between prothetic and metathetic continua. The data we have collected is not in a form which lends itself to giving concrete testimony as to where our scales of frequency and amount fall on this line. Furthermore, even if they are prothetic dimensions, it's still not clear how to resolve Steven's disagreement with others over which technique leads to equal interval scales.

Two levels of topic importance (important and unimportant) were compared across three groups (college students, MBA students, and high school students). Using the untransformed raw data, the multivariate *F* (1.79) representing the comparison of scale values assigned the 39 expressions of frequency when these were imbedded in important and unimportant contexts, respectively, was statistically significant ($df = 78/55$, $p < .05$). Univariate *F*s comparing important with unimportant topics for the words "rather seldom" (4.19), "frequently if not always" (3.97), "very seldom" (4.72) and "a great many times" (4.57) were also statistically significant ($df = 1/132$, $p < .05$).[7]

TABLE 1

Means and Standard Deviations of Expressions of Frequency

Expressions of frequency	$\bar{X}$	*SD*
Always	58.01	3.52
Continually	50.16	3.18
Constantly	49.70	3.31
Frequently if not always	45.24	3.06
Very often	42.45	3.08
A great deal of the time	41.37	3.08
Very frequently	40.02	3.33
A great many times	39.28	3.09
Usually	39.18	3.13
Often	37.64	3.09
Frequently	36.07	3.29
Quite often	35.39	3.93
Rather frequently	34.44	3.18
Commonly	32.97	3.14
Fairly often	32.64	3.30
Fairly many times	30.65	3.00
Sometimes	19.42	2.86
Some of the time	18.01	3.01
To some degree	15.52	2.92
Now and then	15.19	3.04
Occasionally	14.92	3.06
Once in a while	10.22	2.89
Not often	7.78	2.55
Not very often	7.23	2.56
Fairly infrequently	6.99	2.72
Infrequently	6.47	2.61
Rather seldom	6.42	2.66
Very seldom	4.72	2.64
Rarely	4.56	2.23
Very infrequently	4.54	2.47
Seldom if ever	3.69	2.42
Hardly at all	3.47	2.38
Hardly ever	3.34	2.23
Very rarely	2.99	2.11
Almost never	2.63	2.10
Seldom	.33	2.60
None of the time	.17	1.49
Not at all	.15	1.53
Never	.08	1.41

TABLE 2

Means and Standard Deviations of Expressions of Amount

Expressions of amount	$\bar{X}$	*SD*
All	66.12	4.32
An exhaustive amount of	59.27	4.71
Almost entirely	57.61	5.76
Completely	57.35	5.90
An extraordinary amount of	54.46	4.38
Almost completely	51.38	4.71
An extremely abundant amount of	48.89	4.52
An extreme amount of	48.20	3.59
A great amount of	41.56	2.74
A great deal of	41.36	2.83
Very much	40.59	2.94
A full amount of	40.50	3.79
A lot of	37.10	2.77
Much	35.14	2.64
Quite a bit of	34.24	2.71
A good bit of	32.65	3.13
A considerable amount of	31.44	2.90
Pretty much	30.04	2.87
Fairly much	27.70	2.42
An ample amount of	26.22	2.81
An adequate amount of	24.07	2.60
A moderate amount of	21.80	3.42
Some	18.63	2.92
To some extent	13.42	2.95
To some degree	13.10	2.72
Somewhat	11.75	3.03
A limited amount of	9.57	2.85
A little	7.81	2.50
A small amount of	7.51	2.52
Comparatively little	7.22	2.71
A little bit of	7.20	2.67
Not much	7.02	2.70
A small degree of	5.27	2.50
Very little	5.21	2.45
A slight amount of	5.09	2.58
A meager amount of	4.28	2.66
A scanty amount of	3.68	2.41
A minimum amount of	3.64	2.71
A trifling amount of	3.13	2.59
Scarcely any	2.98	2.20
A trivial amount of	2.85	2.60
An insignificant amount of	2.48	2.11
Hardly any	2.28	2.20
None	.15	1.65

---

[7] Calculation of multivariate degrees of freedom for repeated measures MANOVA: Let K = number of measures on each subject; N_t = total number of subjects; P = number of concomitant vectors (covariates);

The raw data were then transformed in two ways (as suggested by Stevens, 1966, 1971). First, because different subjects might have chosen different anchor values for the referent concepts (e.g., one subject might assign the value [illegible]0 to "sometimes", while to another "sometimes" might mean 100), the data were statistically transformed so that each person's anchor value was the same, namely, 50. To do this, each subject's original anchor value was divided into 50, and all other values were multiplied by the result. The second transformation involved the use of logarithms of the raw data.

Both transformations yielded identical results. Neither the multivariate *F*s nor any of the univariate *F*s were significant. The log-transformed data were then analyzed by subsample. Again there were no significant differences in the mean scale values assigned to expressions of frequency when these expressions were imbedded in important as opposed to unimportant contexts. This conclusion was the same for the MBA students, college students, and high school students that constituted our sample.

A second analysis considered the effect of topic importance and group membership with regard to the 44 expressions of amount in a 2 × 3 repeated measures MANOVA. In contrast to the frequency analysis, neither the multivariate *F* nor any of the univariate *F*s was significant in the analysis of the raw data. The analysis performed on the logarithms of the raw data revealed that neither the multivariate F of .84 ($df = 88/41$) nor any of the univariate *F*s was statistically significant. In addition, there were no significant differences when the data were transformed so that each subject had the same anchor value.

DISCUSSION

The main hypothesis of the study, that judges assign different scale values to expressions of frequency and amount as a function of topic importance, was generally not supported. Analysis of the raw data of the frequency expressions, however, did yield significant differences as a function of topic importance. This result is easily explained, however, because different subjects might have chosen different anchor values for the referent concepts. When either of the appropriate transforms was made (Stevens, 1966, 1971), namely, equal anchor values for all subjects or the use of logarithms of the raw data, these apparent differences disappeared.

Although this hypothesis was rejected in this particular experimental instance, it should be pointed out that acceptance of the null hypothesis (i.e., no significant differences in scale values assigned expressions of frequency and amount as a function of topic importance) does not necessarily prove it to be so. The present results do not provide a sufficient condition for concluding that a single list, though possible, is equally applicable for important and unimportant issues. However, the onus is on the skeptic to unearth evidence to reject the null hypothesis and to state a more plausible one. Nevertheless, researchers should be aware of these cautions when using the data from Tables 1 and 2.

Tables 3 and 4 present, respectively, 4-point through 9-point scales of expressions of frequency and amount as well as the percentage overlap in distribution between adjacent scale points. The expressions chosen to represent the various points on each scale are statistically optimal in the sense that their mean scale values came closest to the exact mathematical values necessary to establish a scale of any given ratio. Also the observed variances about their respective means are small. The means and standard deviations of all the expressions have been presented in Tables 1 and 2 in the hope that the complete lists can more adequately serve particular needs of individual investigators.

Tilton's (1937) overlap statistic, 0, was computed in order to provide an index of the amount of separation present between adjacent scale points. It is clear from an examination of Tables 3 and 4 that as scale fineness increases (i.e., as the number of scale

---

and let T = the portion of experimental space being partialed out.

$$\text{Multivariate } df = \frac{K}{N_t - K - P - T}$$

For example, in the frequency analysis:

$$\frac{K}{N_t - K - P - T} = \frac{39 \times 2}{134 - 78 - 0 - 1} = \frac{78}{55}.$$

TABLE 3

STATISTICALLY OPTIMAL SCALES OF FREQUENCY (WITH PERCENTAGE OF OVERLAP BETWEEN SCALE POINTS)

No. points in scale					
9	8	7	6	5	4
8 Always 24%	7 Always 24%	6 Always 24%	5 Always 5%	4 Always 2%	3 Always <1%
7 Continually 21%	6 Continually 21%	5 Constantly 4%	4 Frequently, if not always 8%	3 Very often 12%	2 Often <1%
6 Very often 24%	5 Very often 13%	4 Often 25%	3 Quite often 1%	2 Fairly many times <1%	1 Sometimes <1%
5 Quite often 42%	4 Rather frequently 2.5%	3 Fairly many times 6%	2 Sometimes 10%	1 Occasionally <1%	0 Never
4 Fairly many times 6%	3 Sometimes 45%	2 Sometimes 10%	1 Once in a while 2%	0 Never	
3 Sometimes 45%	2 Now and then 16%	1 Once in a while 2%	0 Never		
2 Occasionally 16%	1 Not often 7%	0 Never			
1 Not very often 7%	0 Never				
0 Never					

Note. Each percentage shown represents the overlap in distribution between the accompanying entry's scale point and the scale point directly below that entry. For example, there is 24% overlap in response distributions between response at Point 7 and response at Point 8 in the 9-point scale.

points increases) so does the percentage of overlap between the distribution of adjacent scale points. As the scale becomes more coarse (fewer scale points) there is correspondingly less overlap between distributions.

CONCLUSION

We have demonstrated, at least, for three samples of judges at different age, educational, and occupational levels, and for important and unimportant issues, that it is possible to fix quantitative meanings that are associated with verbal judgments of frequency and amount. Furthermore, we have specified expressions of frequency and amount that bear integer relations with each other for 4-point to 9-point scales. Also, we have been able to specify the percentage of overlap of judgments between adjacent points on a given scale.

What we provide in this study are a relatively stable set of positive numbers, beginning at zero, for summarizing the increases associated with modifying expressions of frequency and amount. The results should be useful to scale developers. Scales using the expressions listed should be more comparable quantitatively. If the judgment "sometimes" is set at 19, the judgment "always" has a reliable numerical equivalent of 58 in people's minds. Many summary operations, obviously, are easier to perform using such numbers, as opposed to the words themselves, particularly if the numbers are in a ratio scale with an absolute zero.

Another interesting use of the point scales

TABLE 4

STATISTICALLY OPTIMAL SCALES OF AMOUNT (WITH PERCENTAGE OF OVERLAP BETWEEN SCALE POINTS)

No. points in scale					
9	8	7	6	5	4
8 All 44%	7 All 39%	6 All 18%	5 All 10%	4 All 2%	3 All <1%
7 An exhaustive amount of 18%	6 Almost entirely 31%	5 An extraordinary amount of 7%	4 Almost completely 16%	3 An extreme amount of 3%	2 A great amount of <1%
6 An extreme amount of 29%	5 An extreme amount of 8%	4 A great amount of 17%	3 Very much 2%	2 Quite a bit of <1%	1 A moderate amount of <1%
5 A great deal of 20%	4 A lot of 7%	3 Quite a bit of 5%	2 Fairly much <1%	1 Some <1%	0 None
4 Quite a bit of 5%	3 Fairly much 9%	2 A moderate amount of 12%	1 To some degree <1%	0 None	
3 An adequate amount of 32%	2 Some 9%	1 Somewhat 2%	0 None		
2 Some 5%	1 A limited amount of 4%	0 None			
1 A little 6%	0 None				
0 None					

Note. Each percentage shown represents the overlap in distribution between the accompanying entry's scale point and the scale point directly below that entry. For example, there is 44% overlap in response distributions between response at Point 7 and response at Point 8 for the 9-point scale.

and their numerical equivalents would be to locate the verbal expression most closely matching a mean result that fell between two scale points, for example, suppose the 5-point scale of frequency (Table 3) had been used in a study of some substantive issue, and the mean result for a group of respondents came to 3.5, halfway between "always" and "very often." Table 1 indicates that the expression "continually" (50.16) lies just about halfway between "always" (58.01) and "very often" (42.45). Therefore, "continually" would represent the mean position of the respondents.

Serious errors of labeling alternatives in previously developed scales can also be located. For instance, the Leader Behavior Description Questionnaire (LBDQ) is a popular behavior checklist which uses a 5-point scale ranging from *virtually always* to *virtually never*, with the scale midpoint labeled "occasionally." Note in Table 3 that "occasionally" is one whole interval away from the midpoint of the 5-point scale based on magnitude estimation.

More precision can be obtained for linguistic translations. Just how equivalent are the French "toujours", Spanish "siempre" and English "always"? For cross-language contracts, treaties and agreements, the possibility arises of using the universal language of mathematics to locate zones of disagreement between what were purported to be the same statements in different languages.

It has been generally understood that the number of anchor points does not make much difference in most measurement properties of a scale; that is, whether one uses, four points or eight points does not affect the internal consistency, stability, predictive or concurrent validity of the scale (Bendig, 1954; Komorita, 1963; Matell & Jacoby, 1971, 1972). Our study, however, demonstrates one effect that has been previously overlooked. The percentage of overlap in adjacent judgment increases as the number of anchor points increases. When we move from three to nine points, we pay the price for the increase in scale fineness. The overlap between adjacent point judgments can run as high as 44%.

REFERENCES

Bass, B. M. How to succeed in business according to business students and managers. *Journal of Applied Psychology,* 1968, **52**, 254–262.

Bendig, A. W. Reliability and the number of rating scale categories. *Journal of Applied Psychology,* 1954, **38**, 38–40.

Cliff, N. Adverbs as multipliers. *Psychological Review,* 1959, **66**, 27–44.

Hakel, M. D. How often is often? Comment. *American Psychologist,* 1958, **23**, 533–534.

Hemphill, J. K., & Coons, A. E. Leader behavior description. Columbus: Ohio State University, Personnel Research Board, 1950.

Komorita, S. S. Attitude content, intensity, and the neutral point on a Likert scale. *Journal of Social Psychology,* 1963, **61**, [illegible]–334.

Matell, M. S., & Jacoby, J. Is there an optimal number of Likert-scale items? Study I: Reliability and validity. *Educational and Psychological Measurement,* 1971, **31**, 657–[illegible].

Matell, M. S., & Jacoby, J. Is there an optimal number of alternatives for Likert-scale items? Effects of testing time and scale properties. *Journal of Applied Psychology,* 1972, **56**, 506–509.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. *The measurement of meaning.* Urbana: University of Illinois Press, 1957.

Sheppard, D. The adequacy of everyday quantitative expressions as measurements of qualities. *British Journal of Psychology,* 1954, **14**, 40–50.

Simpson, R. H. The specific meanings of certain terms indicating differing degrees of frequency. *Quarterly Journal of Speech,* 1944, **30**, 328–330.

Stevens, S. S. A metric for social consensus. *Science,* 1966, **151**, 530–541.

Stevens, S. S. Issues in psychophysical measurement. *Psychological Review,* 1971, **78**, 426–450.

Thurstone, L. L. Psychophysical analysis. *American Journal of Psychology,* 1927, **38**, 368–389.

Tilton, J. W. The measurement of overlapping. *Journal of Educational Psychology,* 1937, **28**, 656–662.

Uhrbrock, R. S. Standardization of 724 rating scale statements. *Personnel Psychology,* 1950, **2**, 285–316.

(Received April 16, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 321–326

MEASURING ORGANIZATIONAL COMMUNICATION[1]

KARLENE H. ROBERTS[2] AND CHARLES A. O'REILLY III

University of California, Berkeley

Past research on organizational communication has not included systematic development of a measurement method that can be used to compare dimensions of communication within and across organizations. This research reports our initial attempts to develop such instrumentation. Seven samples, with a total number of over 1,200 respondents, were used to develop a 35-item questionnaire measuring 16 facets of communication.

Communication has been the subject of considerable attention in organizational research. However, investigators often attach diverse labels to seemingly similar communication variables and similar labels to what appear to be different variables (Porter & Roberts, in press; Roberts, O'Reilly, Bretton, & Porter, in press). There has been no systematic development of instrumentation to measure communication variables in organizations. The present research begins to develop a standardized questionnaire that can be used to compare communication across organizations.

A review of the organizational literature reveals that among the communication facets most frequently discussed and investigated are directionality of information flow (e.g., Graves, 1972), accuracy and distortion of information (Read, 1962; Wilensky, 1967), modalities used in transmitting information (Lawler, Porter, & Tenenbaum, 1968), and gatekeeping of information (Davis, 1968; O'Reilly & Roberts, in press; Rosen & Tesser, 1970). The communication literature not directly concerned with human behavior in organizations suggests additional aspects of communication that should be given greater attention in behavioral organizational research. These include overload (Porat & Haas, 1969) and satisfaction with one's communication (Lawler, Porter, & Tenenbaum, 1968; Lawson, 1965). Desire to interact with others is suggested by the interpersonal literature to be an important progenitor of actual interaction behavior (Rogers & Bhowmik, 1970). Finally, three noncommunication variables—trust in superior, perceived influence of superior, and mobility aspirations of the respondent—have been repeatedly shown to influence individual communication in organizations (e.g., Cohen, 1958; O'Reilly & Roberts, in press; Read, 1962).

Because communication is a process rather than a static variable it is one of the most difficult organizational variables to measure. Behavioral observations are often impossible to make and self-report forms are usually phrased in such a way that they take snapshots of process variables. The attempt in this study was to develop a measurement device that allows respondents to summarize their own behavior over time. Aggregation of individual responses should minimize many of the inherent individual response errors.

METHOD

Based on a literature review, 189 Likert-type items were created to assess respondent perceptions of the communication dimensions previously indicated. Communication purpose, content, importance, and speed, which are also frequently mentioned in the organizational literature, were considered in initial instrument development. They were later discarded because of the difficulty respondents had answering items relevant to them. Items were also developed to measure desire for interaction and to assess three noncommunication variables: trust in superior, perceived influence of the superior, and respondent mobility aspirations (Likert, 1967; Read, 1962). While the number of researchers who have measured the latter three dimensions is extensive, only Read (1962) reports intercorrelations of items supposedly assessing the same facets.

The 189 items in the pool were administered to 70 graduate students, all of whom had prior work

---

[1] This study was partially supported by Public Health Service Grant MH22054-01 and Office of Naval Research Grant N000314-69-A-0200-1054.

[2] Requests for reprints should be sent to Karlene H. Roberts, School of Business Administration, University of California, Berkeley, 350 Barrows Hall, Berkeley, California 94720.

TABLE 1

VARIMAX FACTOR LOADINGS FOR ITEMS ASSESSING TRUST IN SUPERIOR, PERCEIVED INFLUENCE OF SUPERIOR, AND SUBORDINATE'S MOBILITY ASPIRATIONS FOR A SAMPLE OF 95 MILITARY SUBJECTS

Item	Varimax factor loadings ($n = 95$)		
	1	2	3
1. Trust	*.68*	.22	.13
2. Trust	*.76*	.08	.41
3. Trust	*.79*	.23	.01
4. Perceived influence	.28	.30	*.81*
5. Perceived influence	.14	.16	*.87*
6. Mobility aspirations	.12	*.90*	.21
7. Mobility aspirations	.23	*.88*	.20
Proportion of common factor variance	*.26*	*.26*	*.24*

experience. These students were asked to respond to the questions based on their experience in some past or current job, considering their communication behavior in a typical work week. Items were retained that had high intercorrelations with other items ostensibly measuring the same facets and low intercorrelations with the rest of the item pool. A 60-item questionnaire resulted and was further tested.

This version of the questionnaire (51 communication and 9 noncommunication items) was administered to 86 mental health workers, at all job levels, in five health-care-delivery teams working in a large medical center. Their responses were subjected to a V-type cluster analysis to further reduce the number of questionnaire items. This technique does not seek orthogonal item groups. Instead, using an objective index of collinearity (P^2) and the correlation matrix among variables, the pattern of correlations is compared, and subsets or clusters of variables are identified that have collinear correlation profiles. Each cluster of variables may be thought of as occupying generally the same space, that is, are measuring the same construct (Tryon & Bailey, 1970). Two single communication items were included with the clusters (one measuring overload and the other measuring general satisfaction with communication) after it was determined that these questions behaved appropriately with regard to the rest. The question concerned with satisfaction is a modification of the GM faces scale (Kunin, 1955). When this question was related to the clusters it showed relationships to each cluster without loading heavily on any one, suggesting that satisfaction with communication in an organization is perceived to be partly related to (*a*) interaction, (*b*) accuracy, (*c*) directionality of information flow, etc.

Conceptually, trust, influence, and mobility should be more independent of one another than are the communication items. Consequently these nine noncommunication items were subjected first to a principal-components analysis with initial communality estimates in the diagonal provided by squared multiple correlations (R^2) and then a varimax rotation. Three separate factors emerged. These items were included in a test battery completed by 95 officers and enlisted men in a United States military unit. The results are reported in Table 1.

Based on these analyses the total instrument was reduced to 35 items which were then scored on 10 multi-item indexes and 6 single questions. As previously mentioned, questions concerned with overload, satisfaction, and modality use are reported as individual items.

RESULTS

Reliability

The 35-item version of the instrument was administered twice to 24 graduate business students (all with previous job experience) and to 42 managers for test–retest reliability. In both cases the administrations were three weeks apart. The test–retest reliabilities for the 10 indexes and 6 individual items and a sample question for each dimension are reported in Table 2; 9 of the 10 indexes have reasonable test–retest reliabilities. However, the small sample sizes involved should be kept in mind.

In addition to the mental health workers, the 35-item version of the questionnaire was administered to the military sample previously referenced, 54 emergency room and outpatient hospital personnel, 52 managers from various firms, 179 respondents at various job levels in six United Kingdom locations of a large financial institution, 148 subjects from six comparable United States offices of the same organization, and 579 personnel in another military organization. Differences among the means and standard deviations across the seven samples ($n = 1{,}218$) reveal that within-group variance is less than between-group variance. This suggests that it is feasible to discriminate among organizations using these communication indexes.

Maximum, minimum, and median Cronbach alphas (a measure of the internal consistency of the items measuring a dimension) are reported in Table 3 for each of the seven samples on the 10 multi-item indexes. The alphas appear relatively consistent for all except the last index.

TABLE 2

COMMUNICATION DIMENSIONS AND TEST–RETEST RELIABILITIES

Index	Definer	No. items	Test–retest correlation	
			Students ($n = 24$)	Managers ($n = 42$)
1. Trust	How free do you feel to discuss with your immediate superior the problems and difficulties you have in your job without jeopardizing your position or having it "held against" you later?	3	.71	.69
2. Influence	In general, how much do you feel that your immediate superior can do to further your career in this organization?	2	.49	.62
3. Mobility	How important is it for you to progress upward in your present organization?	2	.77	.80
4. Desire for interaction	How desirable do you feel it is in your organization to be in contact frequently with others at the same job level?	3	.66	.76
5. Directionality—upward	While working, what percentage of the time do you spend in contact with superiors?	3	.84	.87
6. Directionality—downward	While working, what percentage of the time do you spend in contact with subordinates?	3	.82	.83
7. Directionality—lateral	While working, what percentage of the time do you spend in contact with others at the same job level?	3	.82	.76
8. Accuracy	When receiving information from the sources listed below (superior, subordinate, peers), how accurate would you estimate it usually is?	3	.82	.61
9. Summarization	When transmitting information to your immediate superiors, how often do you summarize by emphasizing aspects that are important and minimizing those aspects that are unimportant?	3	.35	.37
10. Gatekeeping	Of the total amount of information you receive at work, how much do you pass on to your immediate superior?	3	.83	.78
11. Overload	Do you ever feel that you receive more information than you can efficiently use?	1	.46	.69
12. Satisfaction	Put a check under the face that expresses how you feel about *communication* in general, including the amount of information you receive, contacts with your superiors and others, the accuracy of information available, etc.?	1	.60	.73
Modalities	Of the total time you engage in communications while on the job, about what percentage of the time do you use the following methods:			
13. Written	Written	1	.44	.63
14. Face-to-face	Face-to-face	1	.86	.41
15. Telephone	Telephone	1	.39	.58
16. Other	Other	1	—	.83

Note. A revision of the questionnaire including additional items and indexes is available on request. Items are scored on 7-point scales, except for indexes 5, 6, and 7 which use 10-point scales.

Validity

Objective criterion measures might be devised for comparison with each of the communication facets discussed in this study. However, difficulty in obtaining objective criterion measures and their potential inappropriateness once obtained warrant consideration of more indirect validity assessments: construct, face, and a very limited attempt at convergent-discriminant validity.

First, evidence for the construct validity of the 10 dimensions is provided by the results of the cluster analyses previously discussed and increases confidence that even similar appearing facets such as trust and influence, or interaction and desire for interaction, can be

TABLE 3

MAXIMUM, MINIMUM, AND MEDIAN CRONBACH ALPHAS FOR THE 10 MULTI-ITEM INDEXES ACROSS SEVEN SAMPLES ($n = 1{,}218$)

Index	Cronbach alpha		
	Median	Maximum	Minimum
1. Trust	.68	.86	.62
2. Influence	.69	.77	.53
3. Mobility	.82	.92	.74
4. Desire for interaction	.68	.74	.29
5. Directionality—upward	.65	.76	.21
6. Directionality—downward	.84	.88	.41
7. Directionality—lateral	.71	.77	.41
8. Accuracy	.62	.71	.37
9. Summarization	.73	.88	.16
10. Gatekeeping	.53	.64	.16

successfully differentiated from each other. The intercorrelations among indexes from a representative sample, shown in Table 4, confirm the notion that communication dimensions are differentiable.

Second, individual questionnaire items appear to have face validity. During the item development respondents indicated which questions were confusing or not differentiable in their minds. Thus, questions attempting to measure task versus social information, speed of transmission, and the purpose of the communication were eliminated.

Third, in the spirit of convergent-discriminant validity, intercorrelations of nine communication and three noncommunication dimensions with other similar and dissimilar measures are presented in Table 5 for the first military sample. The other measures include assessments of overall job satisfaction (Kunin, 1955), supervisor's considerate leadership style (Stogdill & Coons, 1957), organizational competence and flexibility (Campbell & Beaty, 1971), organizational commitment (Porter & Smith, 1970), and respondent's age and tenure.

Examination of Table 5 reveals some of the expected convergence and discrimination. Education and tenure may be surrogates for rank in military units. Thus, it was hypothesized that only communication facets that change as a result of rank are related to edu-

TABLE 4

INTERCORRELATIONS AMONG THE COMMUNICATION INDEXES FOR U.S. BANK EMPLOYEES ($n = 148$)

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Trust	—															
2. Influence	42***	—														
3. Mobility	−07	04	—													
4. Desire for interaction	21*	09	06	—												
5. Directionality—upward	−12	−10	−07	07	—											
6. Directionality—downward	−03	−03	−12	−04	−26**	—										
7. Directionality—lateral	06	−01	09	−04	−30***	−50***	—									
8. Accuracy	20*	26**	−05	29***	06	−05	−08	—								
9. Summarization	12	14	22**	20*	−14	−02	06	14	—							
10. Gatekeeping	04	11	−05	−37***	−04	−09	20*	−27***	−14	—						
11. Overload	−05	01	12	−17*	00	−04	06	03	−11	15	—					
12. Satisfaction	53***	33***	−09	04	−05	00	−15	35***	02	−09	17*	—				
13. Modality—written	03	−07	−03	15	39***	04	−13	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	—			
14. Modality—face-to-face	−05	09	07	−14	−18*	−08	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	—		
15. Modality—telephone	−09	−09	−07	−01	06	23**	−14	10	[illegible]	[illegible]	[illegible]	[illegible]	20*	[illegible]	—	
16. Modality—other	12	15	−09	07	02	−11	08	−04	07	−05	27**	12	−08	−25**	04	—

Note. All scales scored so that a high score equals a high value for that measure. Decimal places are omitted.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

TABLE 5

INTERCORRELATIONS OF NINE COMMUNCIATION AND THREE NONCOMMUNICATION INDEXES WITH OTHER INSTRUMENTS FOR 95 MILITARY RESPONDENTS

Index	Overall job satisfaction (Kunin, 1955)	Leadership-consideration (Stogdill & Coons, 1957)	Organizational competence and flexibility (Campbell & Beaty, 1971)	Organizational Commitment (Porter & Smith, 1970)	Respondent job tenure	Respondent formal education
1. Trust	.38*	.61*	.31*	.25*	−.12	.20
2. Influence	.35*	.36*	.31*	.50*	−.29*	.13
3. Mobility	.53*	.40*	.31*	.57*	−.10	.19
4. Desire for interaction	.46*	.21*	.26*	.37*	−.04	.31*
5. Directionality of contact—upward	−.22*	.04	−.09	−.23*	−.20	−.11
6. Directionality of contact—downward	.33*	.12	.17	.34*	.49*	−.05
7. Directionality of contact—lateral	−.06	.04	−.01	−.04	−.35*	.16
8. Accuracy	.29*	.11	.26*	.21*	−.01	.18
9. Summarization	.27*	.19	.15	.16	.01	.14
10. Amount of information passed	.32*	.27*	.13	.20	.14	.09
11. Overload	−.22*	.01	.08	−.05	−.14	−.17
12. Satisfaction with communication	.37*	.34*	.32*	.33	−.04	−.01

Note. All indexes scored so that a high score represents high values for the measure.
* $p < .05$.

cation and tenure. Of the 18 correlation coefficients involving education and tenure, only the three most logical ones are significant.

Other relationships are also of the type expected. Trust in superior, for example, is related to the respondent's overall job satisfaction, perceived considerate leadership style, perceived organizational competence and flexibilty, and commitment to the organization, but not to tenure and education. The amount of information an individual passes is significantly related to his overall job satisfaction and to his perception of his superior as being considerate, but not to other variables. Desire for interaction also shows appropriate relationships with responses to other perceptions about the organization. The negative relationship between information overload and overall job satisfaction and the positive correlation between communication satisfaction and considerate leadership are also understandable. Finally, the large number of significant relationships among overall job satisfaction and communication and noncommunication indexes seem logical. In sum, the results presented in this study are in no way definitive, but provide a crude approximation of convergent-discriminant validity and increase confidence in the indexes.

Further examples of convergence and discrimination are obtainable by comparing perceptions about communication by people in jobs dependent on information transmission (e.g., dispatchers and operations officers in the military units) with those of people occupying jobs not as sensitive to information transmission (e.g., mechanics). The results of these comparisons (not reported in this study) suggest that communication facets measured by this questionnaire can, in general, discriminate perceptions about communications for occupants of jobs dependent on information from those not so dependent. The former are more discriminating among the communication dimensions (they think more frequently about them).

DISCUSSION

The questionnaire can be considered from three perspectives: comparing organizations, work groups, or homogeneous groups of individuals. Our data suggest organizations can be differentiated on the communication dimensions. This implies that organizations might be taxonomically described in communication terms, for example, a taxonomy based on the openness and accuracy of communication. Work groups can also be contrasted similarly. Groups of individuals (e.g., workers and first-line supervisors) may also display unique communication behaviors. For even greater differentiation of communication problems, practitioners might find it more beneficial to use individual items rather than communication indexes.

The complexities of human behavior in organizations, including communications, should be better understood when they are more rigorously related to other organizational variables. For example, if there is a strong relationship between the degree to which organizational members perceive that the information transmitted at work is accurate and their performance, it should be advantageous for organizations to devote more time to correcting problems of accuracy. This suggests the obvious need to relate communication measures to performance criteria. Determining the relationships of communication variables to objective and behavioral criteria should aid in developing organizational communication models.

Obviously much work remains to be done in developing a standardized measure that can describe various facets of communication across a variety of organizations. The data reported in the present study are exploratory and an initial attempt in this direction. The value of an easily completed descriptive measure of organizational communication seems obvious. If the case often made by organizational theorists is correct—that "good communication" makes a difference—then an understanding of what is "good communication" and its correlates should increase our knowledge of organizational behavior.

REFERENCES

Campbell, J. P., & Beaty, E. Organizational climate: Its measurement and relationship to work group performance. Paper presented at the meeting of the American Psychological Association, Washington, D.C., September 1971.

Cohen, A. Upward communication in experimentally created hierarchies. *Sociological Review,* 1958, **15,** 342–351.

Davis, K. Success of chain-of-command oral communication in a manufacturing management group. *Academy of Management Journal,* 1968, **11,** 379–387.

Graves, D. Reported communication ratios and informal status in managerial work groups. *Human Relations,* 1972, **25,** 159–170.

Kunin, T. The construction of a new type of attitude measure. *Personnel Psychology,* 1955, **8,** 65–78.

Lawler, E., Porter, L., & Tenenbaum, A. Manager's attitudes toward interaction episodes. *Journal of Applied Psychology,* 1968, **52,** 432–439.

Lawson, E. Change in communication nets, performance, and morale. *Human Relations,* 1965, **18,** 139–147.

Likert, R. *The human organization.* New York: McGraw-Hill, 1967.

O'Reilly, C., & Roberts, K. Information filtration in organizations: Three experiments. *Organizational Behavior and Human Performance,* in press.

Porat, A., & Haas, J. Information effects on decision-making. *Behavioral Science,* 1969, **14,** 98–104.

Porter, L. W., & Roberts, K. Organizational communication. In M. Dunnette (Ed.), *Handbook of industrial and organizational psychology.* Chicago: Rand-McNally, in press.

Porter, L. W., & Smith, F. E. The etiology of organizational commitment: A longitudinal study of the initial stages of employee-organization reactions. Unpublished manuscript, University of California, Irvine, 1970.

Read, W. Upward communication in industrial hierarchies. *Human Relations,* 1962, **15,** 3–16.

Roberts, K., O'Reilly, C., Bretton, G., & Porter, L. Organizational theory and organizational communication: A communication failure? In L. W. Porter & K. H. Roberts (Ed.), *Communication in organizations.* Middlesex, England: Penguin, in press. (*Human Relations,* in press.)

Rogers, E., & Bhowmik, D. Homophily-heterophily: Relational concepts for communication research. *Public Opinion Quarterly,* 1970, **34,** 523–538.

Rosen, S., & Tesser, A. On reluctance to communicate undesirable information: The MUM effect. *Sociometry,* 1970, **33,** 253–263.

Stogdill, R., & Coons, A. *Leader behavior: Its description and measurement.* Columbus: Ohio State University, Bureau of Business Research, 1957.

Tryon, R. C., & Bailey, D. E. *Cluster analysis.* New York: McGraw-Hill, 1970.

Wilensky, H. *Organizational intelligence.* New York: Free Press, 1967.

(Received April 30, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 327-331

FACTORS INFLUENCING DISCIPLINARY JUDGMENTS

BENSON ROSEN[1] AND THOMAS H. JERDEE

Graduate School of Business Administration, University of North Carolina

The influence of individual and organizational factors on disciplinary judgment was studied by means of a disciplinary exercise administered to 222 undergraduate business students. In four separate discipline cases, the severity of organizational consequences for a rule infraction and various characteristics of the offenders were manipulated. When the same rule violation resulted in severe compared to mild organizational consequences, greater personal responsibility was attributed to the offender and harsher disciplinary measures were recommended. Offender characteristics such as status, talent, performance, and compensation equity also influenced disciplinary judgments. Results are discussed in terms of current theories of attribution and in terms of practical considerations regarding fair and consistent administration of discipline.

Managerial judgments regarding the severity of particular disciplinary problems and the determination of appropriate disciplinary actions have important organizational implications. An offender's willingness to change his behavior and conform to company policy in the future may depend to a large extent on his perceptions that he has been treated fairly. Moreover, the reactions of other employees to various disciplinary actions may have long-term consequences for organizational morale and future employee behavior. Accordingly, systematic study of factors that influence managerial disciplinary judgments is needed.

Research by Boise (1965) examined employee characteristics that have a bearing on disciplinary action. Boise surveyed Philadelphia city employees and found that a worker's value to his department influenced his supervisor's choice of penalty for rule infraction. According to Boise, supervisors were hesitant to impose penalties on a subordinate when his skills were in short supply.

Kipnis and his associates (Goodstadt & Kipnis, 1970; Kipnis & Cosentino, 1969; Kipnis, Silverman, & Copeland, 1973) investigated the relationship between the supervisor's emotional state and his choice of disciplinary actions. These investigators showed that negative emotional states resulting from uncooperative or hostile subordinate behaviors were associated with more coercive disciplinary measures.

Although there are very few investigations specifically concerned with disciplinary judgments, it seems reasonable to assume that theory and research into attribution of causality and responsibility (Jones & Davis, 1965; Kelley, 1967) is relevant to an understanding of the judgmental processes of managers faced with disciplinary decisions. Attribution theory asserts that greater responsibility will be attributed to an actor, rather than to circumstances beyond the actor's control, in situations where the consequences of the actor's behavior are more severe (Kelley, 1973).

Kelley's (1971) discussion of the processes underlying moral evaluation is also relevant to understanding judgments of responsibility. Kelley argues that judgments of what "works" or what produces favorable outcomes tends to influence judgments of what is morally right. Studies by England (1967) indicate that this pragmatic evaluative framework is especially characteristic of American managers.

The purpose of the present research is to assess the influence of the organizational harm resulting from a rule violation and of various individual characteristics on judgments regarding the severity of disciplinary problems and the appropriateness of various disciplinary actions. Specifically, we predicted that subjects would attribute greater responsibility to an actor and recommend more severe punishment when the unacceptable behavior results in more serious organizational consequences. Also investigated was the hypothesis that an employee's value to the organization (depicted in terms of talent, past performance, and status) and his psychological health influences judgments of the seriousness of his offense and severity of punishment. Finally, it was hypothesized that underpaid employees are treated less harshly than are equitably paid employees.

---

[1] Requests for reprints should be sent to Benson Rosen, Graduate School of Business Administration, University of North Carolina, Chapel Hill, North Carolina 27514.

METHOD

Subjects

The subjects were 222 undergraduate business students at the University of North Carolina.

Procedure

Experimental materials were presented in the form of a discipline exercise. Subjects were asked to assume the role of a personnel manager for a large manufacturer and to consider a series of incidents involving employees who had committed a variety of rule infractions. Each case was written in four or more versions to manipulate the independent variables. From the subjects' perspective, however, the manipulations were completely unobtrusive, since each subject received only one version of each incident and was unaware that other versions existed. Assignments of subjects to treatment conditions were completely random.

The discipline incidents and the experimental variables of interest were as follows:

1. *Illegally using the company car: Extent of Damage × Status of Employee.* Four versions of this incident were written to depict an employee (described as the vice president or janitor) who borrowed the company car for personal use under false pretenses. The car was subsequently involved in an accident described as very minor or as completely demolishing the car.

2. *Taking hard drugs: Department Performance × Employee's Talent.* Four versions of this incident were written to depict a designer (described as marginally talented or exceptionally talented) who often worked under the influence of hard drugs and encouraged fellow employees to use drugs. Departmental performance was described as either high or low.

3. *Extended lunch hour: Consequences for Manager × Employee's Psychological Adjustment.* This incident depicted an administrative assistant (described as well adjusted or suffering from family problems) who frequently extended his lunch hour. The consequences of the extended lunch resulted in failure to keep up with the work or in no noticeable decrement in productivity.

4. *Expense account padding: Salesman's Salary × Salesman's Performance.* Four versions of this incident were written to describe a salesman who was guilty of padding his expense account. The company's sales commission rates were described as among the highest or among the lowest in the industry; and the salesman's performance was described as either up 10% or down 10% in sales over the previous year.

Dependent Variables

Subjects were asked to determine the most appropriate disciplinary action for each incident. Subjects made decisions from the following alternatives: (0) no action; (1) oral warning with no written record; (2) oral warning noted in the employee's record; (3) written reprimand noted on employee's record; (4) one week suspension without pay; (5) discharge. In addition, subjects rated the offenses on three dimensions: (*a*) seriousness, (*b*) offender's personal responsibility, and (*c*) extent of unethical behavior.

In summary, for each of the four disciplinary incidents, each subject read only one version of the consequences of the rule infraction—mild or severe—and read only one version of the mitigating circumstances—designed to increase or decrease attribution of responsibility to the offender.

RESULTS AND DISCUSSION

Effects of Organizational Consequences of Rule Infractions

It was hypothesized that more severe disciplinary action is recommended for rule infractions that result in severe compared to mild organizational consequences. Table 1 shows mean severity of discipline and evaluations of the offender and the offense for the high- and low-consequence-severity conditions in each of the four incidents.

Results supporting the hypothesis were obtained in each of the four incidents. Significantly greater punishment was recommended when illegal use of the company car resulted in a major accident compared to a minor accident; when illegal use of drugs resulted in lower departmental performance compared to satisfactory departmental performance; when an employee's extended lunch hour caused work to fall behind compared to no decrement in work; and when violations of expense account regulations were associated with decreased sales compared to increased sales. In addition, significant differences were also obtained in the predicted direction for evaluations of the seriousness and ethicalness of the rule infractions.

These findings provide strong and consistent support for the influence of severity of consequences on disciplinary decisions.

TABLE 2

EFFECTS OF CHARACTERISTICS OF THE OFFENDER ON DISCIPLINE SEVERITY AND EVALUATIONS OF OFFENSE AND OFFENDER

Dependent variables	Incident 1			Incident 2			Incident 3		
	Janitor ($n = 109$)	Vice president ($n = 113$)	F	Marginal designer ($n = 104$)	Exceptionally talented designer ($n = 113$)	F	Employee with personal problems ($n = 11$[illegible])	Well-adjusted employee ($n = 104$)	F
Severity of discipline	2.98	1.70	147.02**	2.54	2.05	6.12*	1.43	1.40	ns
Seriousness of offense	3.02	1.80	92.00**	2.73	2.40	5.60*	1.60	1.83	3.88*
Unethicalness of offense	2.92	2.13	34.06**	2.70	2.24	7.75*	1.75	1.92	4.55*
Personal responsibility of offender	3.64	3.22	14.22**	3.38	3.20	ns	2.41	2.90	16.57**

* $p < .05$.
** $p < .01$.

Stronger discipline was advocated when violations of organizational policy resulted in greater harm to the organization. These findings also support Kelley's (1971) theory of moral evaluation and imply that judges lose sight of organizational standards or "ideals" in situations where policy violations have minimal consequences for the organization. These findings held for rather blatant violations such as dishonesty and working under the influence of drugs as well as for a lesser infraction of extending the lunch hour.

Greater responsibility for the offense was attributed to employees in the severe compared to the mild organizational consequences conditions in two of the incidents. These findings are consistent with current theories of attribution processes (Kelley, 1973) and recent research by Phares and Wilson (1972), Walster (1966), and Shaw and Sulzer (1964).

Effects of Characteristics of the Offender

It was hypothesized that an employee's value to the organization, depicted in terms of talent, past performance, and status, influences disciplinary judgments. Table 2 shows the mean severity of discipline and evaluations of the offender and the offense for the various employees described in the disciplinary incidents.

An employee's value to the organization significantly influenced recommendations for disciplinary action in two of the incidents. The janitor was disciplined more severely than was the vice president for the identical rule violation, and the designer of marginal talent was disciplined more severely than was the exceptionally talented designer, for working under the influence of drugs. In addition, evaluations of the seriousness and the ethicalness of the behavior and the attributions of the offender's responsibility were in the predicted direction in both of these disciplinary incidents. Offenses committed by employees with high job status or creative talent were perceived as significantly less serious and significantly less unethical, and the offender was perceived as significantly less responsible for the rule infractions. It appears that high status and exceptional talent were associated with greater leniency for rule infractions. These findings provide additional experimental support for Boise's (1965) survey finding that supervisors are reluctant to discipline "highly valued" employees.

A second offender characteristic examined was the psychological adjustment of the employee. In the present context it was hypothesized that an offender is treated with greater sympathy and indulgence because of family problems partially beyond his control. No significant differences were obtained for discipline severity recommendations for the well-adjusted administrative assistant compared to the administrative assistant with family problems. However, the rule violation was perceived as significantly less serious and less unethical, and significantly less personal responsibility was attributed to the assistant with family problems.

Equity Considerations

The third hypothesis considered the relationship between equity and discipline. Greater deference in terms of severity of punishment was shown to the poorly compensated salesman compared to the well-compensated salesman, and this difference reached statistical significance ($F = 4.27$; $df = 1/218$, $p < .05$). Moreover, the dishonest behavior was perceived as significantly less serious ($F = 5.32$, $df = 1/218$, $p < .05$) and significantly less unethical ($F = 9.71$, $df = 1/218$, $p < .01$) for the underpaid salesman compared to the well-paid salesman. Equity considerations also influenced attributions of responsibility for the offense. The equitably paid salesman was perceived to be more responsible for his behavior than was the under-paid salesman ($F = 32.80$, $df = 1/218$, $p < .01$).

It would be interesting to determine in future research the pervasiveness of the equity considerations. In the present sample, the rule infractions can be viewed as restoring equity. Additional research would be required to determine whether rule infractions unrelated to restoring equity, such as deviations from formal work procedures or violation of safety standards, would also be downgraded for under-paid employees. A related question involves the extent to which negative outcomes other than low pay, such as dehumanizing working conditions, temper judgments of appropriate disciplinary action for various rule infractions.

In addition to providing support for current theories of attribution processes, the present findings have some implications of practical importance to managers. While it is quite reasonable to expect that greater disciplinary action would be taken for very serious rule infractions compared to small violations, our data suggest that some observers overreact to the organizational consequences resulting from an offense. A manager who fails to react to a rule violation because it resulted in little harm to the organization might give the offender and other employees the impression that such violations are sometimes tolerated. This inconsistent approach to resolving disciplinary problems could have detrimental effects on trust and morale of all employees. Similarly, detrimental effects might also be expected when employee status is a major basis of differential treatment. The use of disciplinary indulgence as a means of achieving equity is also fraught with difficulties, since it piles one form of inequity atop another.

REFERENCES

BOISE, W. B. Supervisors' attitudes toward disciplinary actions. *Personnel Administration*, 1965, **28**, 24–27.

ENGLAND, G. W. *Personal value systems of American managers*. *Academy of Management Journal*, 1967, **10**, 53–68.

GOODSTADT, B., & KIPNIS, D. Situational influences on the use of power. *Journal of Applied Psychology*, 1970, **54**, 201–207.

JONES, E. E., & DAVIS, K. E. From acts to dispositions. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology*, Vol. 2. New York: Academic Press, 1965.

KELLEY, H. H. Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska Symposium on Motivation*, 1967. Lincoln: University of Nebraska Press, 1967.

KELLEY, H. H. Moral evaluation. *American Psychologist*, 1971, **26**, 293–300.

KELLEY, H. H. The processes of causal attribution. *American Psychologist*, 1973, **28**, 107–128.

KIPNIS, D., & COSENTINO, J. Use of leadership powers in industry. *Journal of Applied Psychology*, 1969, **53**, 460–466.

KIPNIS, D., SILVERMAN, A., & COPELAND, C. Effects of emotional arousal on the use of supervised coercion with black and union employees. *Journal of Applied Psychology*, 1973, **57**, 38–43.

PHARES, E. J., & WILSON, K. C. Responsibility attribution: Role of outcome severity, situational ambiguity, and internal-external control. *Journal of Personality*, 1972, **40**, 392–406.

SHAW, M. E., & SULZER, J. L. An empirical test of Heider's levels in attribution of responsibility. *Journal of Abnormal and Social Psychology*, 1964, **69**, 39–46.

WALSTER, E. Assignment of responsibility for an accident. *Journal of Personality and Social Psychology*, 1966, **3**, 73–79.

(Received April 23, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 332–336

EFFECTS OF INCENTIVE, FEEDBACK, AND MANNER OF PRESENTING THE FEEDBACK ON LEADER BEHAVIOR

RICHARD P. BUTLER[1]
U.S. Military Academy

CABOT L. JAFFEE
Florida Technological Institute

The purpose of this study was to determine if the content of a leader's verbal behavior could be altered by manipulation of the type and manner of feedback and incentive. Ninety-six groups of four people, with one leader per group, served as subjects in a 2 × 2 × 2 design. During a case discussion, the experimenter varied the three independent variables and recorded the content of the leader's verbal behavior by means of the Bales category system. It was found that: (*a*) positive feedback made the leaders more task oriented, whereas negative feedback increased negative social–emotional behavior; (*b*) the incentive decreased the number of questions asked; and (*c*) the manner of presenting the feedback (publicly or privately) produced no changes in verbal behavior when analyzed as a main effect, but significant interactions were found.

In trying to identify the causes of leadership-attempting behavior in small groups, many recent studies have focused on situational rather than personal determinants of this type of behavior. Jaffee (1968), in developing his empirically based theory concerning the determinants of leadership-attempting behavior, as measured by verbal behavior, exemplifies this approach. He concluded that the more readily accessible method of modifying leadership-attempting behavior of individuals within a group may be the manipulation of the characteristics of the situation rather than the long-term "personality" of the individual. Jaffee further noted that three of the prime causes involved in whether an individual attempts to lead and the amount of his leadership-attempting are: the availability of both incentive and feedback, and the manner in which the feedback is given. However, it was not known if these variables affected the content or kind of verbal behavior that an individual emits when attempting leadership behavior. It was the purpose of this study to assess the effects of the aforementioned variables on the content or quality of a leader's verbal behavior. Specifically, the variables of incentive, feedback, and the manner in which the feedback was given (either privately or publicly) were manipulated, and the resulting kinds of verbal behavior were recorded.

METHOD

Subjects

The subjects were 384 male volunteers who were taking an introductory psychology course. In all, 96 groups of 4 were observed.

Experimental Setting and Apparatus

The setting was an 8 × 12-foot room separated from an observation room by a one-way mirror. A table large enough to accommodate four persons was placed in this room. On the table, facing the subject designated as the leader, was placed a box through which feedback was given. The box had a total of four lights, two of which were located on the inner side and visible only to the leader, and two of which were placed atop the box and in full view of all the subjects. Each pair consisted of one red light and one green light. A microphone was positioned in the middle of the table and was connected to an amplifier with earphones so that the subjects' discussion could be monitored in the adjoining room. The switches for operating the lights of the leader's feedback box were located in this adjoining room.

Experimental Design

A 2 × 2 × 2 complete factorial design with equal replications per cell was employed. Twelve groups of four served in each of the eight randomly assigned treatment conditions. An explanation of the factors follows.

Incentive. For one half of the 96 groups the leaders were told that they would receive $5 if they did an effective job in leading the ensuing discussion. No mention of the $5 incentive was made to the leaders of the other 48 groups.

---

[1] Requests for reprints should be sent to Richard P. Butler, Office of the Director of Institutional Research, U.S. Military Academy, West Point, New York 10996.

Feedback. For one half of the groups the leaders received primarily positive feedback via the flashing of lights, that is, eight green lights indicating that the leaders were doing well and two red ones meaning that they were not doing well. For the remaining 48 groups the leaders were given mainly negative feedback or eight red lights and two green lights. For each leader the feedback lights were presented in a random order subject to the ratio constraints previously mentioned.

Manner of presenting the feedback. Forty-eight leaders received their feedback publicly, that is, the red and the green lights on top of their feedback box were flashed, thus enabling the other group members to also see how well their leader was doing. For the remaining 48 leaders the feedback was given privately so that only the leader himself knew how well he was doing.

The Bales category system (Bales, 1950) was used to record the qualitative or content verbal behavior of the leaders.

Procedure

The four members of each group were read general instructions by the experimenter. In essence, the instructions stated that the purpose of the experiment was to observe and study how a small group goes about solving two separate problems that often occur in the everyday working world. The first case problem concerned a salesman who constantly violated company rules but nevertheless was high in sales. The group's task was to come to some kind of solution to the case in the allotted time of five minutes. The sole purpose of this first discussion was to identify the duration of speech for each subject so that the individual who spoke the most could be appointed as the leader for the second case discussion. It has been shown that duration of speech by individuals in small groups is highly correlated with a group member being selected as the leader of the group (Simpkins & West, 1966). The experimenter recorded duration of speech from behind the one-way mirror in the control room. At the end of the five minutes he reentered the experimental room, appointed the individual with the longest duration of speech as the leader, and read instructions to the second case problem. This case involved the distribution of funds to the various departments of a small town for the coming year. Again, the task of the group was to find a solution. Instructions to the leader varied according to which treatment he had been assigned. During the course of the ensuing discussion the experimenter observed the leader's behavior, recorded his verbal remarks on the Bales category system, and delivered feedback to him every two and one-half minutes. When the subjects agreed on a solution to the case or when the 30-minute time limit had elapsed, the experimenter entered the room and ended the discussion.

To assess the experimenter's reliability in scoring the Bales category system, a second observer, who had experience in the use of category systems, simultaneously observed five groups and recorded with the experimenter.

RESULTS

For each of the 12 Bales's categories the frequency of behaviors considered within its domain was recorded for each of the 96 leaders. However, for purposes of data analyses, the 12 categories were not treated separately, but, following a suggestion by Bales (1950), were collapsed into four major areas: (*a*) Categories 1–3 (positive social–emotional acts), (*b*) 4–6 (attempted answers), (*c*) 7–9 (questions), and (*d*) 10–12 (negative social–emotional acts).

To assess scorer reliability, Pearson product-moment correlations were computed between the experimenter's scores and those of the other observer for five groups. All coefficients for the four major areas were .90 or above. Thus, high reliabilities as reported by Bales (1950) were obtained in this study as well.

By far, the largest average proportion (.70) of the total number of acts falling within each of the major areas for the 12 leaders in each of the eight treatment combinations was in the attempted answer area. Positive social–emotional acts and the question area were in a virtual tie for second place with average proportions of .12 and .13, respectively. Negative social–emotional behavior accounted for an average proportion of only .05 of the acts. Examination of the proportions, using Bartlett's test for homogeneity of variance, indicated that heterogeneity of variance across treatments was present. To remove this difficulty so that statistical assumptions could be met, the arcsin transformation was employed. Similar analyses on the transformed data showed that the transformation succeeded in obtaining homogeneity of variance.

To analyze the effects of incentive versus no incentive, positive versus negative feedback, and public versus private presentation of this feedback on each of the four major Bales's categories, four 2 × 2 × 2 analyses of variance were made. Table 1 presents the results when the *F* values were significant. If an *F* value was not significant, it does not appear in Table 1. Thus, for Category A and

TABLE 1

ANALYSES OF VARIANCE OF PROPORTIONS ON EACH OF BALES' MAJOR CATEGORIES

Behavior category	*df*	*MS*	*F*
C. Questions			
Incentive	1	.15	4.62*
D. Negative social-emotional acts			
Feedback	1	.47	16.79***
Incentive × Manner of Feedback	1	.23	8.04**
A+D. Social-emotional acts			
Feedback	1	.18	6.84**
Incentive × Manner of Feedback	1	.19	7.35**
B+C. Tasks related units			
Feedback	1	.26	7.94**

Note. Proportions were transformed by $x' = 2 \arcsin \sqrt{x}$. Only significant *F* values are shown. All others were nonsignificant.

* $p < .05$.
** $p < .01$.
*** $p < .001$.

Category B there were no significant main or interactive effects; and therefore their analyses do not appear in the table. For Category C, the only significant effect was for the incentive versus no incentive treatment ($F = 4.62$, $df = 1/88$, $p < .05$). Inspection of the means showed that the *no* incentive condition elicited significantly more questions than did the $5 incentive treatment. It can be conjectured then that without the pressure of a $5 incentive the leaders were more prone to ask for the opinions, suggestions, and information of other group members.

For Category D, a highly significant difference was obtained between the primarily negative and primarily positive feedback conditions ($F = 16.79$, $df = 1/88$, $p < .001$). Examination of the means revealed that when the mostly negative feedback treatment was administered there was a strong increase in the number of behaviors falling within the negative social–emotional area. It appears then, that when negative feedback is common, the leaders tend to become more disagreeable, tense, and antagonistic, since these are behaviors listed under Category D. It should be noted that virtually all the groups needed the full 30 minutes to come to a solution to the problem. Thus, the number of feedbacks presented to each leader was 10. Another significant effect was the first-order interaction between the incentive conditions and the manner of presenting the feedback, that is, privately or publicly ($F = 8.04$, $df = 1/88$, $p < .01$). Inspection of the means showed that when no incentive was present with public presentation of the feedback, the greatest amount of negative social–emotional behavior was exhibited. In contrast to this, the smallest amount was produced when no incentive was combined with private feedback.

A second method of analyzing the Bales category system is to see if the treatment combinations produced any differences in the proportions of acts directed to the task area (B + C) and to the social–emotional area (A + D). Two additional analyses of variance were performed on these summary indices. The results of the analysis of the A + D combination also appear in Table 1. Two significant *F* ratios were obtained, the first was for the feedback variable ($F = 6.84$, $df = 1/88$, $p < .01$). Under negative feedback, a significant increase in the number of acts in the social–emotional area was elicited. The second significant *F* ratio ($F = 7.35$, $df = 1/88$, $p < .01$) was the first-order interaction between incentive and the manner of administering the feedback. In this study, the least amount of social–emotional behavior occurred when no incentive and private feedback were employed; the greatest amount occurred under conditions of no incentive-public feedback.

The analysis of variance for the B + C combination (Table 1) indicated a significant effect for the feedback conditions ($F = 7.94$, $df = 1/88$, $p < .01$). The inference from looking at the means was that when primarily positive feedback was given, the leaders tended to be more task oriented.

DISCUSSION

The purpose of the present study was to see if, by contriving the environment in which the leader of a small group performed, the content of his verbal behavior could be altered. It was found that it was possible to modify some, but not all, types of verbal behavior. The fact that some types were subject to modification is important in that it adds a new dimension to the modification of leadership behavior in small groups, namely that not only can the sheer quantity of speech be altered, as has been shown by several other researchers (see Jaffee, 1968), but also that the content of leader verbal behavior can be modified as well.

Specifically, of the three independent variables that were manipulated to create different environments, it was found that the feedback factor had the greatest effect of any single factor. Positive feedback made the leaders more task oriented, whereas negative feedback increased their social–emotional remarks. Apparently, positive feedback had the effect of producing leaders who were more willing to delve into the main issue at hand and more willing to give and ask for information, suggestions, and opinions. Negative feedback, on the other hand, caused leaders to be more disagreeable and to show tension and antagonism at a higher rate. Thus, the results of this study indicate that in a production-oriented organization, positive feedback is to be preferred to negative feedback, and negative feedback might have very little to offer if no specific suggestions for changing one's behavior are given. In fact, Deutsch (1949), House (1966), and Schutz (1955) all have shown that conflict and antagonistic behavior often lead to a decrease in production.

Content verbal changes in the social–emotional area were also noticeable as a result of the first-order interaction between the incentive factor and the manner of presenting the feedback. However, the significant interaction pertaining to A + D was probably caused by the difference in the proportion of acts falling in Category D, since no differences were noticed in Category A. Therefore interpretation is limited to the interaction that produced significant differences in Category D. It is important to note that neither factor alone was powerful enough to produce significant changes. This means that to produce changes, it might be necessary to administer both of them simultaneously. It is also important to consider the combination that elicited the low and high levels of negative social–emotional behavior. Lack of incentive led to fewer remarks under conditions of private feedback than under public feedback, whereas incentive resulted in just the opposite effects. When analyzing the results that were obtained when the $5 incentive was paired with public or private feedback, the fact that the public feedback combination resulted in fewer negative remarks indicates that it produced the more socially desirable behavior.

A third area in which changes in verbal content were produced occurred in Category C, questions. It was found that incentive was able to decrease the number of times that a leader asked for information, opinions, and suggestions from his subordinates. This finding is somewhat consistent with Reilly and Jaffee (1970), who found that under conditions of incentive, individuals were less willing to be "led" by others. Apparently, then, leaders prefer to take things into their own hands, that is, to become less democratic when there is a chance at financial remuneration.

The two remaining Bales's Categories, A and B, were not subject to modification by any of the independent variables. It may be that these categories are extremely difficult to modify under any laboratory conditions or that the independent variables were just irrelevant for them.

In summary, it appears that the variables involved in Jaffee's (1968) theory concerning the amount of leadership-attempting behavior have at least partial relevance to the modification of the content of a leader's verbal behavior. More specifically, the feedback variable seems to possess some generality to both the amount and content of a leader's verbal behavior, the incentive factor somewhat less, and the manner of presenting the feedback, publicly or privately, none at all to the content area, except when combined with incentive in certain circumstances.

REFERENCES

BALES, R. F. *Interaction process analysis: A method for the study of small groups*. Cambridge, Mass.: Addison-Wesley, 1950.

DEUTSCH, M. An experimental study of the effects of cooperation and competition upon group process. *Human Relations*, 1949, 2, 199–231.

HOUSE, W. C. Effects of group cohesiveness on organization performance. *Personnel Psychology,* 1966, **45**, 28–33.

JAFFEE, C. L. Leadership attempting: Why and when. *Psychological Reports,* 1968, **23**, 939–946.

REILLY, R., & JAFFEE, C. L. Influence of some task-irrelevant factors on leader selection. *Psychological Record,* 1970, **20**, 535–539

SCHUTZ, W. C. What [illegible] groups productive? *Human Relations,* [illegible]–465.

SIMPKINS, L., & WEST, J. [illegible] Reinforcement of duration of talking in trial groups. *Psychological Reports,* 1966, **18**, 231–236.

(Received March [illegible], 1973)

(*Continued from page 263*)

A Utility Analysis of a Weighted Application Blank Designed to Predict Turnover for Clerical Employees. Raymond Lee (Hennepin County Personnel Office, 605 Fourth Avenue South, Minneapolis, Minnesota 55415) and Jerome M. Booth.

Prediction of Scuba Training Performance. Robert J. Biersner (Department of the Navy, Pensacola, Florida 32508) and David H. Ryman.

Some Results with the German Form of the Position Analysis Questionnaire (PAQ). Ekkehart Frieling (Universitat Augsburg Memminger Strasse 14, 8900 Augsburg, Germany), R. Lindberg, and W. [illegible]heiser.

Effects of Head Restriction on Drivers' Eye Movement and Errors in Simulated Dangerous Situations. John A. Allen, Jr. (Highway Safety Research Center, University of North Carolina, Chapel Hill, North Carolina 27514), Stephen R. Schroeder, and Patricia G. Ball.

Cross-Sample Stability of the Supervisory Behavior Description Questionnaire. Henry P. Sims, Jr. (Graduate School of Business, Indiana University, Bloomington, Indiana 47401) and Andrew D. Szilagyi.

Effects of the Shorter Workweek on Selected Satisfaction and Performance Measures. John M. Ivancevich (College of Business and Economics, University of Kentucky, Lexington, Kentucky 40506).

Disparagement by a Subordinate as an Influence on the Use of Power. Eugene M. Foder (Department of Psychology, Clarkson College of Technology, Potsdam, New York 13676).

Comment on Locke's Satisfiers and Dissatisfiers Among White-Collar and Blue-Collar Employees. Robert G. Hadley (Rehabilitation Counseling Program, California State University, 5151 State University Drive, Los Angeles, California 90032).

Journal of Applied Psychology
1974, Vol. 59, No. 3, 337–338

PARTICIPATIVE MANAGEMENT IN THE COLLEGE CLASSROOM

RALPH H. KILMANN[1]

University of Pittsburgh

It was hypothesized that students encouraged to provide information and decision-making efforts for classroom procedures of importance to them give more favorable ratings to various aspects of the instructor's teaching effectiveness and to their perceptions of what they learned than do students not given this encouragement and opportunity. Two graduate classes designed the course within the latitudes of administrative constraints. As controls, two graduate classes were presented the identical course design the following semester as dictated by the instructor. Results indicated reasonable support for the hypothesis.

Several studies have investigated the differences between student-centered versus instructor-centered teaching vis à vis learning and motivation (for an extensive review, see Anderson, 1959). In general, with objective examinations used as an indicator of learning, the two types of teaching do not indicate any consistent differences. However, for "critical thinking" (i.e., essay exams and problem analysis) and for motivational criteria, student-centered teaching seems more effective (McKeachie, 1969). A systematic understanding of *why* and *how* student-centered methods contribute to the goal of education that goes beyond acquisition of knowledge could lead to useful changes in classroom activities. This study assumes that the theories developed to improve managerial and organization effectiveness are relevant to this understanding of the classroom situation.

Research in industry has found that giving individuals greater influence in decision making increases the effectiveness of their performance and their levels of satisfaction (e.g., Bass & Leavitt, 1963). It seems reasonable to ask if the decision-making process in the classroom can be designed to increase the quality of learning and motivation of students.

Not all activities and circumstances of the classroom, as in any life area, are subject to group participation in decision making. For example, a school administration may expect that the class will meet according to some time schedule, that formal grading will be performed by the instructor, and that there will be some distribution of grades. Even within such constraints, however, there is latitude in the ways the constraints are satisfied. Much of the day-to-day actual classroom activity is not specified by the administration.

The present study hypothesized that when students are encouraged to participate in decision making regarding classroom procedures, they give more favorable ratings to various aspects of the instructor's teaching effectiveness and to their perceptions of what they learned.

METHOD

Subjects

Subjects were graduate business students in four classes (each of about 25 students) of a required course. The same instructor taught all four classes. Two daytime classes, taught in the winter semester, were assigned to the experimental treatment; two night classes in the spring semester were used as controls. Consequently, it was unlikely that students in these two treatments would interact and, therefore, discover differences between the instructor's methods for each class.

Procedure

On the first day of class for the experimental groups, the instructor outlined the constraints on the course. Within broad limits, he stated that the students, if they so desired, could determine the basic design of the course. Basically, the instructor permitted the students a voice in classroom issues they felt important to them.

In the two control classes, the instructor dictated the course design that was identical to that developed by the students in the experimental groups.

---

[1] Requests for reprints should be sent to Ralph H. Kilmann, Graduate School of Business, University of Pittsburgh, Pennsylvania 15213.

One week before the end of the semester, each class was administered a teacher and course evaluation form (University of Pittsburgh's Student Opinion of Teaching Questionnaire) by a student representative. This form contains 22 items concerning the student's perception of what he learned, as well as his reactions to various aspects of the instruction.

For each item (e.g., "The instructor has given me new viewpoints") a rating is made on a 5-point scale (1 = instructor ranks below most of the teachers you have known, 2 = is only fair, 3 = is competent and compares well, 4 = is well above average, 5 = is one of the most outstanding).

RESULTS

Of the 22 variables tested, the experimental subjects ($n = 54$) rated all items higher than did the controls ($n = 45$), except for contribution to learning via lectures. According to the sign test (Siegel, 1956), this overall result is statistically significant well beyond the .001 level. In addition, for eight separate items the difference in favor of higher mean ratings for the experimental group were statistically significant by a two-tailed t test ($p < .05$). These items were the following: the instructor has increased my interest, presents worthwhile material, encourages initiative, has stimulated thinking, shows respect for opinions, and is sensitive to student difficulties; the class discussions contributed to learning, and would recommend the course as an elective.

DISCUSSION

The only variable which was not rated higher for the experimental condition concerned the "contribution to learning" made by the lectures. Since both classes in the control condition were taught after those in the experimental condition, it is possible that the instructor's lecturing technique actually improved with practice for the control group.

A limitation in the design is that other differences between the day and night students may account for these results. However, in the graduate program in business at the University of Pittsburgh, the differences in class composition and motivation did not seem significant. Whether the experimental effect is informational (e.g., students may have a better understanding) or motivational is not possible to establish in the present study. Further, a limitation of the study is that only one instructor participated. Although this helped to control for instructor differences, the results may not be generalizable to the teaching styles of other instructors. The study does suggest that follow-up work on participative management in the classroom may be fruitful.

REFERENCES

ANDERSON, R. C. Learning in discussion: A resume of the authoritarian–democratic studies. *Harvard Educational Review*, 1959, 29, 201–267.

BASS, B. M., & LEAVITT, H. J. Some experiments in planning and operating. *Management Science*, 1963, 9, 574–585.

MCKEACHIE, W. J. *Teaching tips: A guide for the beginning college teacher*. Lexington, Mass.: Heath, 1969.

SIEGEL, S. *Nonparametric statistics*. New York: McGraw-Hill, 1956.

(Received April 9, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 339–343

INTRINSIC, EXTRINSIC, AND RISK PROPENSITY FACTORS IN THE WORK ATTITUDES OF TEACHERS, EDUCATIONAL ADMINISTRATORS, AND BUSINESS MANAGERS[1]

CECIL MISKEL[2]

University of Kansas

Parallel versions of Borgatta, Ford, and Bohrnstedt's Work Components Study Questionnaire were used to compare the job orientations of college-educated employees in educational and business organizations. The sample was composed of 432 teachers, 118 educational administrators, and 192 business managers. Multiple- and single-classification analysis of variance procedures were used to explore relationships and group differences across the attitudinal subscales. The hypothesis that the attitudes of educators differ from the attitudes of business managers toward intrinsic–extrinsic rewards and risk orientations was supported. Finally, a continuum of differing group profiles was proposed.

During school board elections the opinion that schools should be "put on a businesslike basis" becomes particularly evident. One implication of this is that schools should be administered from a competitive efficiency model as are business enterprises. This model seemingly places a high value on risk propensity and intrinsic factors, while placing a low value on hygiene factors. Consequently, the norms regarding risk taking in the decision-making process and the level and type of work incentives potentially are fundamental variables in determining the usable range of administrative models for different organizations.

This potential importance of employee attitudes toward administering organizations and a lack of sufficient comparisons between organizational types invite further investigation. Consequently, the purpose of this study is to compare the orientations of college-educated teachers, educational administrators, and business managers toward risk propensity, extrinsic factors, and intrinsic factors in their jobs.

RATIONALE AND HYPOTHESIS

The conceptual foundation of work attitudes for this study was developed from the works of Herzberg, Mausner, and Snyderman (1959) and Blum (1961). Herzberg found that positive events are dominated by references to intrinsic aspects of the job itself, while negative events are dominated by references to extrinsic aspects of the job situation. Sergiovanni (1967) concluded that the results of replicating Herzberg's study in the educational organization supported the findings in industry.

However, Blum's findings modify these conclusions, showing that the desire for security can be a deciding factor in vocational choice. A person might select a job where elements of security and other hygiene factors are high. Other persons, less concerned with security, choose jobs in which motivators or satisfiers are high. Williams' (1965) findings that low-risk takers are more concerned with extrinsic work characteristics and that high-risk takers are more concerned with intrinsic characteristics support Blum's conclusion.

Pavalko (1971) places the work orientations of educators and business managers at opposite ends of an occupation–profession continuum with the norm in business occupations favoring personal aggressiveness. Krause (1971) supported this position, noting that the ability to handle uncertainty is a

---

[1] This research was supported by a grant from the General Research Office at the University of Kansas.

[2] The author wishes to thank L. E. Heller, M. S. Patton and B. W. Thornton for their assistance in the data collection and analysis procedures.

Requests for reprints should be sent to Cecil Miskel, School of Education, University of Kansas, Bailey Hall, Lawrence, Kansas 66045.

basic career contingency for business managers. Further support is provided by the findings that individuals scoring highest on the Work Components Study Questionnaire subscale measuring competitiveness desirability were perceived as contributing the most to the company (Ford, Borgatta, & Bohrnstedt, 1969). Conversely, the norms in educational organizations appear to place a premium on a more reserved, less aggressive orientation. In a simulated problem-solving situation, Brown (1970) found that business managers have a higher risk propensity and achievement motivation than do educational administrators. Based on these positions, the hypothesis was derived that the attitudes of educators differ significantly from the attitudes of business managers toward Herzberg's motivation factors and Blum's security orientations as measured by parallel forms of the Work Components Study Questionnaire.

METHOD

The instruments used to measure the attitudes were the Work Compenents Study and the Educational Work Components Study questionnaires. Borgatta (1967) developed the Work Components Study to merge and operationalize Herzberg's two-factor theory with Blum's findings regarding the security orientation among workers. Essentially, the items ask the respondents to judge the desirability of jobs with varying amounts of intrinsic factors, extrinsic factors, and intrinsic combined with risk factors.

Miskel and Heller (1973) modified the Borgatta instrument to suit the educational organization by replacing words relating to industrial work situations with words pertaining to an educational work situation. They used three statistical techniques to test the instrument's adaptability to the educational situation. Principal-components, varimax orthogonal, and maxplane oblique *R*-factor analysis procedures were used to determine the factorial stability of the Educational Work Components Study, while Cronbach's (1951) alpha coefficient was calculated to estimate the internal consistency of the items.

The criteria for determining the number of factors were the following: scree test, discontinuity of eigenvalues (Cattell, 1968), interpretability (Rummel, 1970), the findings of earlier research on the WCS (Borgatta et al., 1968), and Kaiser's (1960) eigenvalue of one. In addition, five criteria were used in selecting items: (*a*) minimum factor loadings of .40 after varimax rotation, (*b*) minimum pattern loadings of .40 after maxplane rotation, (*c*) minimum structure loadings as product-moment correlations of the items with the oblique factors (Harmon, 1967) of .40, (*d*) items loading on the factors indicated by the original developers (Borgatta et al., 1968), and (*e*) minimal cross-loading on two or more factors. No a priori preferences were made regarding the importance of these criteria. Consequently, when the criteria were in conflict, a judgment was made as to which made the most overall sense.

The result was parallel measures composed of 56 Likert-type items representing six factors, whose forms are identical except for references to company or school. A description of the six factors based on the work of Ford et al. (1969) and the highest orthogonally loaded item follows.

1. Potential for Personal Challenge and Development (eight items). This factor contains items to measure the desire for creativity and responsibility in the job. The highest factor loading was .63 for the item "I would have a chance to further my formal education."

2. Competitiveness Desirability and Reward of Success (seven items). These items measure whether an individual seeks job situations in which the salary is determined by merit and the competition is keen. The item "salary increases would be a matter of how much effort you put in" had the highest loading of .64.

3. Tolerance for Work Pressure (nine items). This factor contains items measuring attitudes toward situations in which the work load might be excessive. The highest factor loading was .65 for the item "the work might build up pressures on me."

4. Conservative Security (11 items). These items measure the individual's desire for security with well-defined promotion guidelines and job routines. The item "the work would be routine but not hard to do" had the highest factor loading of .75.

5. Willingness to Seek Reward in Spite of Uncertainty versus Avoidance of Uncertainty (10 items). This factor contains items measuring the individual's willingness to do interesting work even though it might be a temporary job. The highest factor loading of .75 was found for the item "there is little permanency of positions."

6. Surround Concern (11 items). These items measure the person's concern with the hygienic aspects of the job. The item "the ventilation is modern" had the highest factor loading of .73.

The intercorrelations and Cronbach's alpha coefficients as estimates of reliability for the factors are presented in Table 1. With the exceptions of Factors 3 and 5 correlating at .55 and Factors 1 and 6 correlating at .39, the six Educational Work Components Study factors are relatively independent. The reliability estimates of the original seven-factor, 66-item Work Components Study have a range from .65 to .85 (Borgatta et al., 1968). The reliability estimates of the six-factor, 56-item Educational Work Components Study Questionnaire compare favorably with the earlier results, ranging from .73 to .83. The reliability estimates of the 56-item Work Components Study with the present sample was somewhat lower with a range from .61 to .84.

The Work Components Study and Educational Work Components Study were administered as self-report forms. The respondents read: "How desirable would YOU consider each of the following items in a job for YOU. A job where. . . ." The items followed, each with a five-category, Likert-type response varying from "Completely undesirable, would never take the job" to "Extremely desirable, would favor the job greatly." The response categories were assigned ascending values of one to five.

Sampling Procedures

Educators. A stratified random sample of 683 teachers and administrators was selected from three school districts with combined staffs totaling 4,200 in a midwestern metropolitan region. The following three proportions were used in selecting subjects: (*a*) of the teaching staffs, 25% in the smallest district and 17.5% in the two larger districts, (*b*) 50% of the administrators, and (*c*) a separate sample of 50 elementary male teachers. Using an original and two follow-up mailings, analyzable returns were received from 118 administrators and 432 teachers for a total of 550 returns or 80.5% of the selected sample.

Business managers. A sample of 265 business managers was randomly selected from seven industrial concerns in the same metropolitan area. The representation from each type is as follows: 140 from computing and communication service companies, 60 from clothing, steel, and tool manufacturing companies, and 65 from insurance and brokerage companies. Using an original and one follow-up mailing, 192 or 72.4% of the questionnaires were returned for analysis.

Data Analysis Procedures

Multiple-classification analysis of variance was used to explore the overall relationships among the variables and groups, while single-classification analysis of variance was used to establish those variables that contributed to group discrimination. Where significant *F* ratios were found using the univariate procedure, the Tukey (*a*) and, in one case, the multiple contrast post hoc procedures (Winer, 1971) were used to compare all pairs of group means. The omega-squared was calculated as outlined by Halderson and Glasnapp (1972) to estimate the strength of the variable relationships.

The analyses were computed for four groups—teachers, principals, central office administrators, and business managers. Since the programs required equal sample sizes and the smallest group had 48 subjects, 48 subjects were randomly selected from each of the remaining groups.

RESULTS

The initial step was using a multivariate analysis of variance procedure with the four employee groups as the independent variable and the six Educational Work Components Study–Work Components Study factors as the dependent variables. In the present context an approximate multivariate *F* statistic of 5.77 ($df = 18/518$, $p \leq .001$) was obtained. Based on this significant *F* statistic, tests for simple main effects became necessary across each of the Educational Work Components Study–Work Components Study subscales. These further analyses are presented and discussed in the following paragraphs.

A summary of the data for the four groups across each of the six Educational Work Components Study–Work Components Study factors comprises Table 2. It shows that the groups differed significantly ($df = 3/188$, $p \leq .05$) across the following subscales: Competitiveness Desirability and Reward of Success, Tolerance for Work Pressure, Conservative Security, and Surround Concern.

Using the aforementioned post hoc procedures to compare all pairs of group means, the following differences were obtained. First,

TABLE 1

INTERCORRELATIONS OF EDUCATIONAL WORK COMPONENTS STUDY FACTORS AND ALPHA COEFFICIENTS FOR EDUCATIONAL WORK COMPONENTS STUDY AND PARALLEL WORK COMPONENTS STUDY FACTORS

Factor	1	2	3	4	5	6
1. PPCD	.80 .69	.13	.13	−.17	.17	.39
2. CDRS		.73 .61	.30	−.01	.15	.24
3. TWP			.79 .78	−.23	.55	−.01
4. CS				.81 .84	−.27	.26
5. WSR					.82 .80	−.12
6. SC						.83 .82

Note. Data for Educational Work Components Study are for 683 educators and for parallel Work Components Study are for 192 business managers. Cronbach's alpha coefficients are in the main diagonal with the Educational Work Components Study alphas above those for the Work Components Study. PPCD = Potential for Personal Challenge and Development; CDRS = Competitiveness Desirability and Reward of Success; TWP = Tolerance for Work Pressure; CS = Conservative Security; WSR = Willingness to Seek Reward in spite of Uncertainty versus Avoidance of Uncertainty; SC = Surround Concern.

TABLE 2

Means, Standard Deviations, and Analysis of Variance Results for the Four Groups on the Educational Work Components Study–Work Components Study Factors

Factor	Teachers		Principals		Administrators		Managers		F	p	ω^2
	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*			
PPCD	4.2	.5	4.0	.4	4.2	.4	4.1	.4	.88	.54	.00
CDRS	3.3	.7	3.4	.6	3.9	1.3	4.0	.4	9.89	.00	.14
TWP	3.0	.6	3.4	.4	3.5	.5	3.5	.4	9.66	.00	.14
CS	2.8	.6	2.9	.6	2.9	.6	2.6	.5	2.75	.04	.03
WSR	2.3	.5	2.4	.5	2.5	1.0	2.6	.6	2.03	.11	.02
SC	4.0	.4	4.0	.3	4.0	.7	3.7	.4	4.88	.003	.06

Note. PPCD = Potential for Personal Challenge and Development; CDRS = Competiveness Desirability and Reward of Success; TWP = Tolerance for Work Pressure; CS = Conservative Security; WSR = Willingness to Seek Reward in spite of Uncertainty versus Avoidance of Uncertainty; SC = Surround Concern.

on the Competitiveness Desirability and Reward of Success subscale, central office administrators and business managers scored significantly higher than principals and teachers. Second, on the Tolerance for Work Pressure factor, teachers scored significantly lower than the other three groups. Third, on the Conservative Security subscale, the Tukey(a) procedure did not indicate any significant differences among the groups even though the mean for the business managers appeared to be much lower (see Table 2). Consequently, the multiple contrast post hoc procedure was used to explore these relationships further. The calculated contrast weights furnished evidence that the business group's mean is significantly lower ($p \leq .01$) than a combination of the educator groups' means on this subscale. Fourth, on the Surround Concern measure, the Tukey(a) procedure revealed that the business group has a significantly lower mean than has each of the three educator groups.

To assist the interpretation of these differences, the average item means and standard deviations by subscale for each group also are presented in Table 2. From these descriptive data, three systematic trends in the relationships become evident. First, the three subscales composed of attitudes toward intrinsic rewards and risk—Competitiveness Desirability and Reward of Success, Tolerance for Work Pressure, and Willingness to Seek Reward in spite of Uncertainty versus the Avoidance of Uncertainty—have means that increase directly from teachers to principals to central office administrators to business managers. Second, the two factors measuring attitudes toward extrinsic job outcomes—Conservative Security and Surround Concern—are approximately the same for the educator groups but markedly lower for the business group. Finally, only on the subscale measuring attitudes toward intrinsic rewards—Potential for Personal Challenge and Development—did all four groups have approximately equal mean scores.

The omega-squared values also are presented in Table 2. These values generally support the F values and explain 2% to 14% of the total attitudinal variance due to group differences.

Discussion

Based on these findings, the hypothesis that educators' attitudes differ from business managers' attitudes toward intrinsic–extrinsic job outcomes and risk orientations was generally supported and now can be elaborated. The educators and business managers are similar in their attitudes toward intrinsic work factors since no differences were found on the Potential for Personal Challenge and Development subscale. Apparently all of the groups equally desire opportunities for creativity and responsibility. However, when a risk dimension is added to the intrinsic orientations—as in the Competitiveness Desirability and Reward of Success, Tolerance for Work Pressure, and Willingness to Seek Reward in spite of Uncertainty factors—significant differences and consistent trends emerge. The business managers had the highest mean scores on these three subscales, while the teachers had the lowest although not significantly lower on Willingness to Seek Reward. The two educational administrator groups placed between the teacher and manager groups, with the central office administrator group appearing higher only on the Competitiveness Desirability and Reward of Success subscale. Finally, when attitudes toward extrinsic rewards are compared in the Conservative Security and Surround Concern factors, the educator groups are similar, but the business manager group is significantly different.

Synthesizing these results yields a conceptual continuum of differing group profiles. The business managers evidence attitudes with high-risk propensity and less concern for extrinsic factors. At the opposite pole, the teachers show low-risk propensity with a high concern for extrinsic factors. The two educational administrator groups, appearing in the middle of the continuum, are similar in their attitudes to teachers in having a high concern for extrinsic factors and security; but when risk is attached to intrinsic factors, they resemble the business managers.

REFERENCES

BLUM, S. H. The desire for security. *Journal of Educational Psychology*, 1961, **52**, 317–321.

BORGATTA, E. F. The work components study: A set of measures for work motivation. *Journal of Psychological Studies*, 1967, **16**, 1–11.

BORGATTA, E. F., FORD, R. N., & BOHRNSTEDT, G. W. The work components study: A revised set of measures for work motivation. *Multivariate Behavioral Research*, 1968, **3**, 403–414.

BROWN, J. S. Risk propensity in decision making: A comparison of business and public school administrators. *Administrative Science Quarterly*, 1970, **15**, 473–481.

CATTELL, R. B. Extracting the correct number of factors in factor analysis. *Educational and Psychological Measurement*, 1968, **18**, 791–837.

CRONBACH, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, **16**, 297–334.

FORD, R. N., BORGATTA, E. F., & BOHRNSTEDT, G. W. Use of the work components study with new college-level employees. *Journal of Applied Psychology*, 1969, **53**, 367–376.

HALDERSON, J. S., & GLASNAPP, D. R. Generalized rules for calculating the magnitude of an effect in factorial and repeated measures ANOVA designs. *American Educational Research Journal*, 1972, **9**, 301–310.

HARMON, H. N. *Modern factor analysis*. (2nd ed.) Chicago: University of Chicago Press, 1967.

HERZBERG, F., MAUSNER, B., & SNYDERMAN, B. B. *The motivation to work*. New York: Wiley, 1959.

KAISER, H. F. The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 1960, **20**, 141–151.

KRAUSE, E. A. *The sociology of occupations*. Boston: Little, Brown, 1971.

MISKEL, C. G., & HELLER, L. E. The educational work components study: An adapted set of measures for work motivation. *Journal of Experimental Education*, 1973, **42**, 45–50.

PAVALKO, R. *Sociology of occupations and professions*. Itasca, Ill.: F. E. Peacock, 1971.

RUMMEL, R. J. *Applied factor analysis*. Evanston, Ill.: Northwestern University Press, 1970.

SERGIOVANNI, T. J. Factors which affect satisfaction and dissatisfaction of teachers. *Journal of Educational Administration*, 1967, **5**, 66–82.

WILLIAMS, L. K. Some correlates of risk taking. *Personnel Psychology*, 1965, **18**, 297–309.

WINER, B. J. *Statistical principles in experimental design*. (2nd ed.) New York: McGraw-Hill, 1971.

(Received May 29, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 344–353

CLUSTER ANALYSES OF INTEREST INVENTORY SCALES AS TESTS OF HOLLAND'S OCCUPATIONAL CLASSIFICATION [1]

DEAN H. NAFZIGER [2] AND SAMUEL T. HELMS

Johns Hopkins University

This study compared Holland's occupational categories with groups of occupations that resulted from the application of McQuitty's iterative intercolumnar correlational analysis to the scales of the Strong Vocational Interest Blank (SVIB), the Minnesota Vocational Interest Inventory (MVII), and the Kuder Occupational Interest Survey (OIS) for men and women. The results indicated that clusters of occupations exist that are internally consistent, and these usually agreed with the groups of occupations in Holland's classification. The hierarchical structure of the clusters followed the hexagonal ordering of Holland's occupational categories suggested in other studies. In addition, the usefulness of all three letters in Holland's occupational classification was supported.

Considerable research has been conducted by Holland and his colleagues in testing and revising a classification of occupations (Holland, 1973; Holland, Viernstein, Kuo, Karweit, & Blum, 1972; Holland, Whitney, Cole, & Richards, 1969). The classification contains six main categories—Realistic (R), Investigative (I), Artistic (A), Social (S), Enterprising (E), and Conventional (C)—and 72 subcategories within the main categories, such as Realistic–Investigative–Enterprising (RIE), Realistic–Investigative–Social (RIS) and so forth. All occupational categories were derived from a single set of coordinating definitions, six scale scores from the Vocational Preference Inventory, VPI, (Holland, 1970).

A hexagonal ordering (see Figure 1) of the six types based on intercorrelations of the VPI scales was proposed by Holland et al. (1969). On the hexagon, the adjacent categories are the most closely related types; the nonadjacent and nonopposite categories are the next most closely related types; and the opposite categories on the diagonal are the least related types. For example, Conventional types are most highly related to Realistic or Enterprising types, moderately related to Investigative or Social types, and least related to Artistic types.

This report investigated the usefulness of the theoretical occupational classification system developed by Holland. The objective was (*a*) to obtain occupational data from sources independent of the information used to develop Holland's classification, (*b*) to group these occupational data using an empirical technique, and (*c*) to compare the results of the empirical grouping with Holland's theoretical grouping. The correspondence between the results of the empirical and theoretical clusters provided a test of the validity of Holland's occupational classification system.

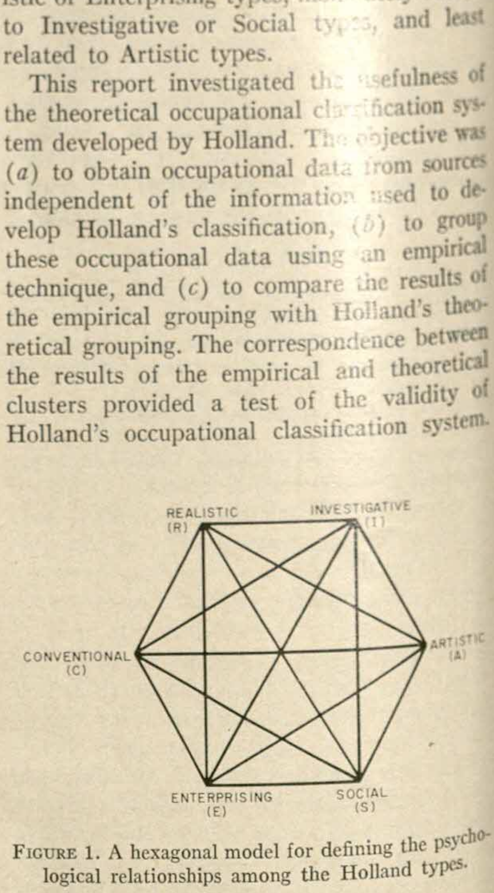


FIGURE 1. A hexagonal model for defining the psychological relationships among the Holland types.

---

[1] This study was conducted under the Careers and Curricula Program at the Center for Social Organization of Schools, Johns Hopkins University, with funding support from the U.S. Office of Education.

[2] Requests for reprints should be sent to Dean H. Nafziger, Center for Social Organization of Schools, Johns Hopkins University, Baltimore, Maryland 21218.

Method

The data used for the analysis of the Strong Vocational Interest Blank (SVIB) were the intercorrelations for the 72 occupational scales of the SVIB (Campbell, 1971, pp. 456–458). Analysis of the Minnesota Vocational Interest Inventory (MVII) was done using the 21 × 21 intercorrelation matrix of the MVII occupational scales found in the manual (Clark & Campbell, 1965). Data used for the analysis of the Kuder Occupational Interest Survey (OIS) were the intercorrelations of the 23 core scales for men and the intercorrelations of the 21 core scales for women (Kuder, 1971, pp. 56–59).

Because this research concerned placing objects or variables into discrete categories using multivariate information, a cluster-analytic technique was appropriate (Borgen & Weiss, 1971). Clustering of the scales was accomplished by using iterative, intercolumnar correlational analysis (McQuitty & Clark, 1968). The clusters that result from this analytical technique are internally consistent types. In the technique, the successive division of the original intercorrelation matrix proceeds first by grouping many objects with relatively few characteristics in common and then by grouping few objects with many characteristics in common. Misclassifications occurring at higher levels are necessarily carried on to lower groupings. It is useful to cross-validate the clusters with other data, but this was not possible in the present study.

To compare the resulting clusters and subclusters to the groupings of the Holland classification, weighted summary codes were obtained for each cluster and subcluster. To accomplish this task, Holland codes, as listed in his classification, were assigned to the occupations in the cluster or subcluster. In the classification, every occupation is assigned a three-letter code, which describes an occupation's characteristics. Taken in order, the three letters of the code indicate the primary, secondary, and tertiary orientation of a job. For example, physician (ISA) is primarily an Investigative occupation, secondly Social, and thirdly Artistic. The empirical derivation of the occupational codes has been summarized by Holland (1973).

After assigning three-letter codes to the occupations in the clusters or subclusters, the number of times a letter appeared in the first, second, or third position of an occupational code was counted and given a weight of 3, 2, or 1, respectively, to account for the relative importance of a letter's position. Summing the weighted values for each letter and rank ordering all letters provided a group summary code for the cluster or subcluster. For large groups, as occurred in the first divisions of the occupations in the cluster analyses, the group summary code concealed the diversity of the individual occupational codes. However, it provided information about the general orientation of even heterogeneous clusters. At lower division points, the group summary code more accurately described the occupations within the cluster.

Results

SVIB Clusters

Figures 2 and 3 summarize the clusters resulting from the analysis of the SVIB.[3] The first major division of the occupational scales yielded two large groupings: Main Cluster 1 and Main Cluster 2. Occupations of Main Cluster 1 may be characterized as scientific, literary, and artistic occupations; those of Main Cluster 2, as business, administrative, and applied science occupations. The group summary code for Main Cluster 1 was I/A S (Investigative/Artistic Social) and for Main Cluster 2 was I/S E (Investigative/Social Enterprising). (Slashes separating letters indicate ties in the summary code scores.)

Main Cluster 1—Professional Occupations

The first primary cluster was a relatively homogeneous group of professional occupations. With respect to the hexagon in Figure 1, this group consisted primarily of occupations coded with permutations of the three adjacent letters IAS, although the R and E dimensions had a small influence. The 28 occupations within Main Cluster 1 subdivided into two groups of 14 each. When Holland codes were assigned to these occupations, one subdivision appeared to be IA (Investigative Artistic) directed while the other appeared to be SA (Social Artistic) directed.

IA directed: Natural scientists, artists, intellectuals. This cluster was comprised primarily of natural scientists, artists, and intellectuals. As a result, most Holland codes of these occupuations had IA, AI, or IS in the first two positions, and the summary code for the cluster was IAS. On the basis of their Holland codes, college professor (SIA) and musician (ASI) were slightly out of place in the IA-directed cluster, as they would have obtained a closer match with their codes in the SA-directed cluster. The IA-directed cluster itself broke into two groups: one group contained artists and intellectuals with

---

[3] Summary figures only are shown throughout this article. The complete cluster tree diagrams that are summarized by Figures 2 through 6 may be obtained on request from the authors at the Center for Social Organization of Schools, Johns Hopkins University, Baltimore, Maryland 21218.

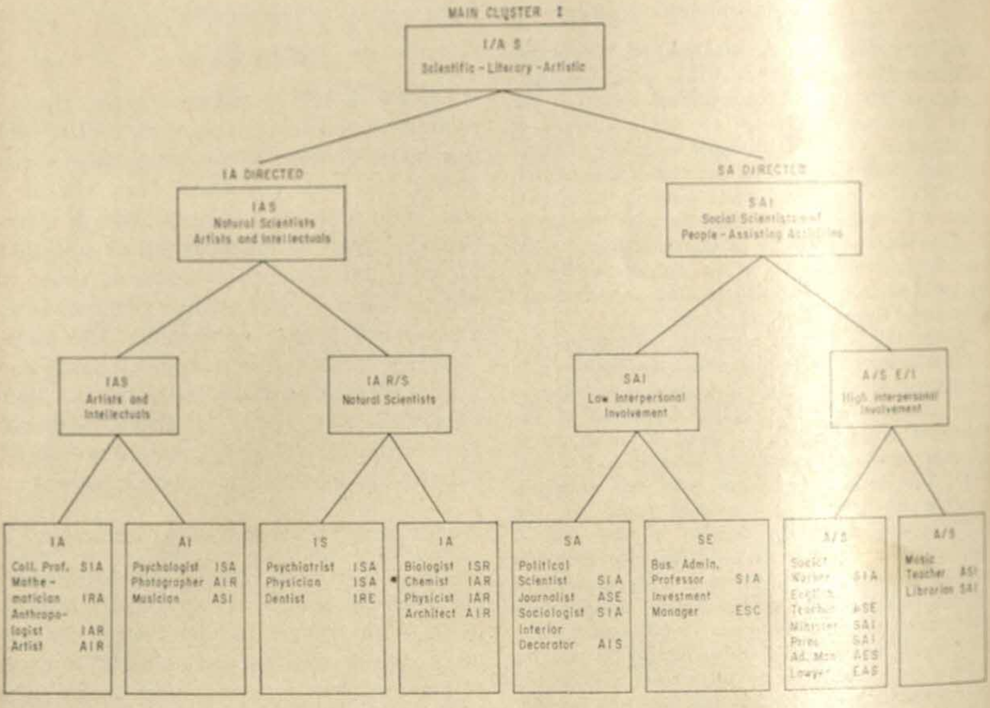


FIGURE 2. Main Cluster I for Strong Vocational Interest Blank (SVIB) occupational scales with abbreviated group summary codes and descriptive labels.

high I and A interests (summary code IAS); the other contained natural scientists with predominantly I characteristics (summary code IA R/S).

SA directed: Working with people. Occupations in this cluster were clearly oriented toward working with people. The demand for a high degree of interpersonal activity was reflected in the strong S orientation of the codes, which mainly had SA, AS, or SI in the first two positions. The summary code for the cluster was SAI. The E category, which is an adjacent hexagonal category to S, occurred frequently in the codes of this cluster. In fact, two predominantly Enterprising occupations—lawyer (EAS) and investment manager (ESC)—fell in this cluster. The second of these occupations was incongruent with the SA-directed cluster.

The SA-directed cluster broke into two subclusters, the basis for division apparently having been the level of personal involvement required. The first subcluster contained occupations that observe human behavior in a detached manner, such as political scientist, journalist, sociologist, and investment manager. The summary code for this subcluster was SAI. Occupations in the second subcluster—for example, social worker, priest, lawyer, and music teacher—demand direct involvement with people. Its summary code, A/S E/I, reflected a relatively more diverse composition.

Main Cluster 2—Residual Occupations

Main Cluster 2, shown in Figure 3, was more heterogeneous than the first large cluster. It appeared to be a residual set of occupations that did not belong with the relatively more homogeneous I/A S occupations in Main Cluster 1.

The two large clusters resulting from the subdivision of Main Cluster 2 were an object-data-oriented cluster and a business-sales-management-oriented cluster. The first of the clusters appeared to be IR directed and the second, SE directed.

IR directed: Data and object manipula-

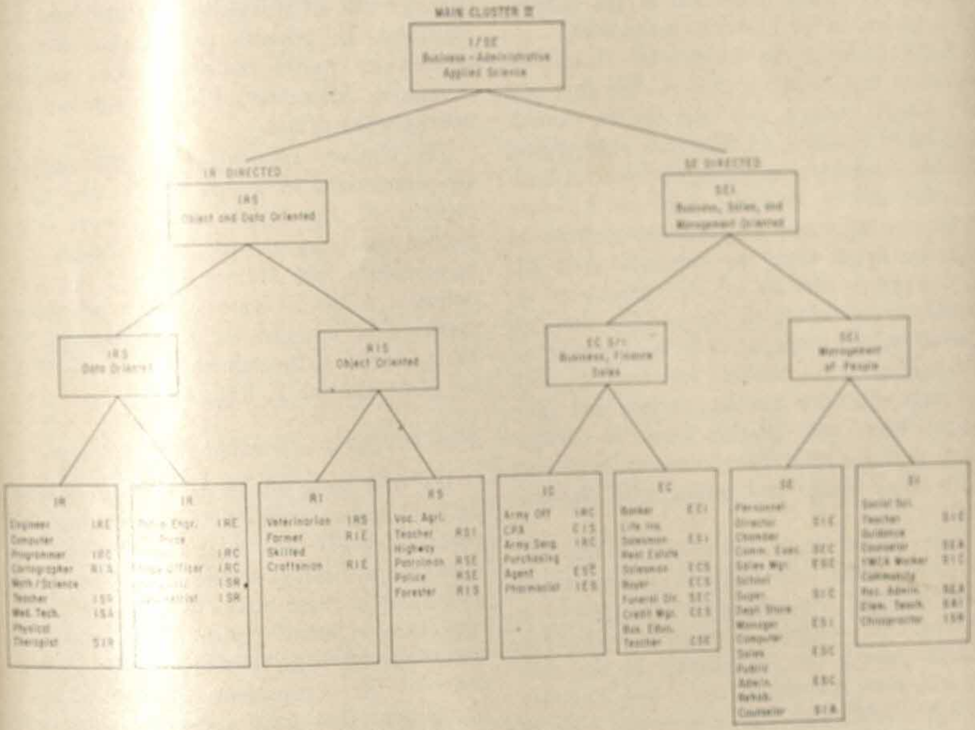


FIGURE 3. Main Cluster II for Strong Vocational Interest Blank (SVIB) occupational scales with abbreviated group summary codes and descriptive labels.

tion. The 18 occupations in this cluster were oriented toward the manipulation of data and objects, which was reflected in their Holland codes. Every occupational code but one began with either R or I, and most of the codes had RI or IR in the first two positions. The summary code of the cluster was IRS. The joint R and I orientation of this cluster was consistent in view of the relatedness of the types. Both R and I types are relatively task oriented and nonsociable, preferring occupational situations with low-interpersonal involvement. That differences do exist between the R and I categories was demonstrated when the IR-directed cluster was further subdivided. The resulting split left one cluster which was directed toward the I vertex of the hexagon and a second cluster which was directed toward the R vertex. Cartographer (RIA) and veterinarian (IRS) were the only two occupations that fell into an unexpected occupational cluster.

It is noteworthy that a subcluster involving RS-coded occupations occurred. R and S are opposite types in Holland's theory as indicated by their positions on the hexagonal configuration of the types. Therefore, occupations combining RS codes should consist of people unique in their relationship to other occupations. That this is indeed the case was reflected by the result of a single cluster containing all RS occupations with one RIS occupation.

SE directed: Business, sales, management. The second group of Main Cluster 2, containing 26 occupations, was the largest and most heterogeneous of the clusters. As a group, the occupations in this cluster had a strong orientation toward business, sales, and management. The summary code for the cluster was SEI. More than any of the other clusters, the occupations in this group required a high level of competence in interpersonal relations. Interpersonal involvement was an important dimension in the cluster, as is typical in all E and S occupations.

Subdivision of the SE-directed cluster resulted in two large subclusters. The first was a business, finance, and sales cluster. Combinations of E and C in the codes were typical in this subcluster, as expected from its business orientation. However, the S and I dimensions also appeared in these occupations as may be noted from the summary code EC S/I. Further division of this subcluster resulted in one group of occupations that appeared to be a pure set of ECS permutations (beginning with banker and ending with business education teacher, finance and sales occupations), and another group of occupations that was difficult to interpret (beginning with Army officer and ending with pharmacist).

The second large subcluster of the SE-directed cluster was composed of occupations that involved the management of people in institutional settings. Consistent with Holland's theory, the occupations in the subcluster were dominated by S and E codes, and the subcluster had a summary code of SEI. Three occupations with somewhat incongruent codes appeared in the cluster: rehabilitation counselor (SIA), elementary teacher (SAI), and chiropractor (ISR).

A further division of this second subcluster resulted in one group of occupations that involved sales and administration and another group that involved contact with people in educational and recreational settings. Most occupations in the sales and administration group had a first-letter code of E. With one exception, the occupations in the educational and recreational subcluster had a first-letter code of S. The Holland codes for these groups were consistent with the general orientation of the occupations in the clusters.

MVII Clusters

The MVII was developed to measure the interest patterns of men in nonprofessional occupations. When Holland occupational codes were assigned to each of 21 occupations in the inventory, 16 were classified as Realistic, two as Social, two as Enterprising, and one as Conventional. In Holland's terminology, then, the MVII measures interests that are characteristic of primarily Realistic type occupations. In general, an analysis of these occupations would be one which examines intragroup differences of interests of a rather homogeneous group.

The clusters resulting from this analysis are summarized in Figure 4. The first major division of the MVII occupational scales yielded two large groups: Main Cluster 1 (9 occupations) and Main Cluster 2 (12 occupations). A people versus object orientation appeared to account for this grouping. For Main Cluster 1, the summary code was RSE; for Main Cluster 2, RIC.

Main Cluster 1—Oriented Toward People

The first large cluster of the MVII occupational scales consisted of occupations that require interpersonal interaction, often in a salesperson-customer relationship. The sort of work done in these occupations is of the light duty, clean hands type. The orientation toward people was reflected in the Holland codes of the occupations. All of the MVII occupations that have codes beginning with S, E, or C were found in Main Cluster 1.

Main Cluster 1, having a summary code of RSE, branched into two subdivisions of three and six occupations each. The summary codes were SR I/C/E for the former, and RES for the latter. This grouping divided those occupations with largely a SR orientation from those occupations with relatively less of S but with more E and C components. That a unique set of characteristics appears to accompany RS codes was again suggested by the separation of warehouseman, the only RS occupation, from the larger RES subcluster.

Main Cluster 2—Orientation Toward Objects

The second large cluster was oriented toward object manipulation. In contrast with Main Cluster 1, this cluster was typified by occupations involving heavy work, manipulation and modification of materials, and getting one's hands dirty. Generally the occupations were blue-collar jobs requiring little close, interpersonal contact, and this was reflected by a paucity of an S dimension in the codes. Without exception the occupations in Main

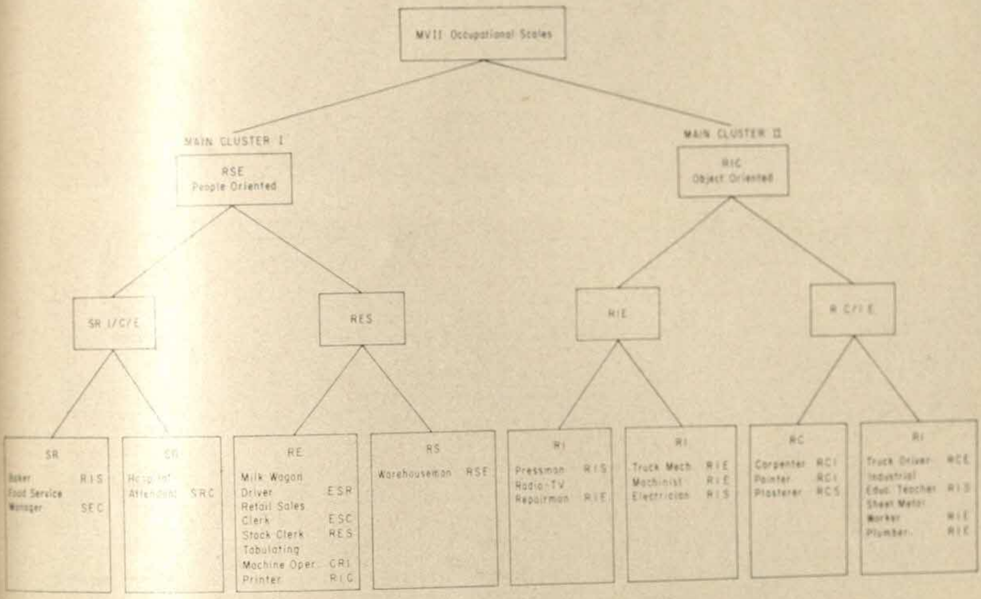


FIGURE 4. Main Clusters I and II for Minnesota Vocational Interest Inventory (MVII) occupational scales with abbreviated group summary codes and descriptive labels.

Cluster 2 had R as the first letter of the Holland code. The summary code for the cluster was RIC.

At the first division of Main Cluster 2 a notably clean split occurred. One subdivision consisted of five occupations that had identical first and second letter codes of RI followed by either S or E as the third letter. The other subdivision contained seven occupations with either RC or RI as the first two letters. The former appeared to be an RI-oriented cluster, and the latter an R C/I-oriented cluster.

The RI-oriented cluster further divided into one group of two and one group of three occupations, each with the same RI summary code. The second subdivision, the R C/I cluster, divided into two groups of three and four occupations each. The division occurred at the second letter (a tie), with RC occupations grouping together and RI occupations grouping among themselves. There was one exception to this: truck driver (RCE).

Kuder DD OIS Clusters for Men

Figure 5 summarizes the clusters resulting from the analysis of the OIS core scales for men. The first major division yielded two groups: Main Cluster 1 (8 occupations) with an RIE summary code, and Main Cluster 2 (15 occupations) with a summary code of ISA. Within Main Cluster 1 were occupations that typically deal with objects and data. Those within Main Cluster 2 may be described as quantitative and life science occupations combined with a subcluster of people and business-oriented occupations. This clustering in terms of object-data versus scientific-social-business was reflected in the Holland group summary codes of RIE and ISA for the two clusters, respectively.

Main Cluster 1—Object, Data and Thing Oriented

The occupational titles for the scales in this first primary cluster mirrored the orientation (of this group) toward manipulation and modification of objects and data. Subdivision of Main Cluster 1 resulted in a group with five occupations and another with three occupations. In the subdivision with three occupations, the engineering occupations and engineering college major remained together and composed an object-data-oriented sub-

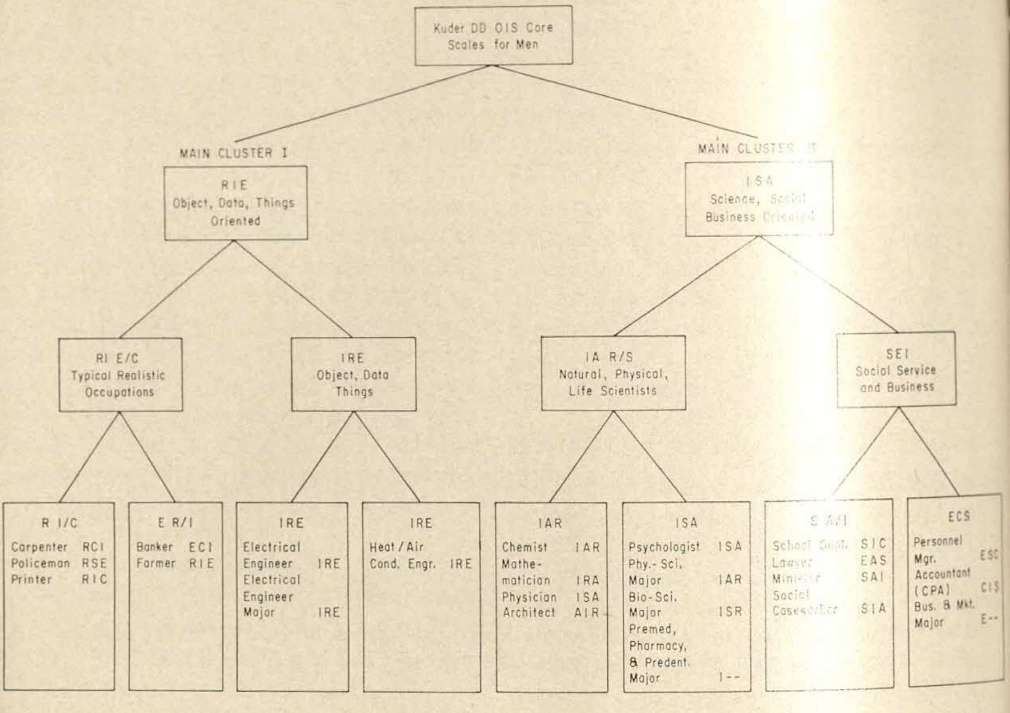


FIGURE 5. Main Clusters I and II for Kuder DD Occupational Interest Survey (OIS) core scales for men with abbreviated group summary codes and descriptive labels.

cluster with an IRE summary code. The occupations that remained together in the other subcluster had an RI summary code and, although not as homogeneous as the group of engineers, seemed congruent with one another. Four of the five occupations required activities, competencies and interests typical of Realistic people. The one exception was banker (ECI).

Main Cluster 2—Science, Social and Business Oriented

Scientific, social, and business occupations comprised this large group, and the ISA summary code for the main cluster reflected these components. Those jobs which are primarily natural, physical, or biological scientists and generally quantitatively oriented remained together in an IA R/S subcluster; while those that are people and business oriented grouped together in an SEI cluster. The summary codes described well the nature of the difference between the occupations in the two subclusters.

The occupations of the first subcluster all had codes that were permutations of I, A, R, and S. With the exception of architect, all began with I. (The I-- code was used as a gross summary for the premedicine, pharmacy, and dentistry major.) The nature of the occupations and their Holland summary code of IA R/S reflected a good deal of congruency. The subcluster further subdivided into two other groups, one seemed to be more quantitively oriented and the other more life science oriented.

The second subcluster, coded SEI, was comprised of social and business occupations. After further division, the four socially oriented occupations remained together while the three business occupations clustered together. The former subcluster had a Holland summary code of S A/I, and the code of the latter was ECS. These codes were quite descriptive of the occupations within the groups. Accountants, usually coded CES, were actually all certified public accountants, hence the CIS code applied to them.

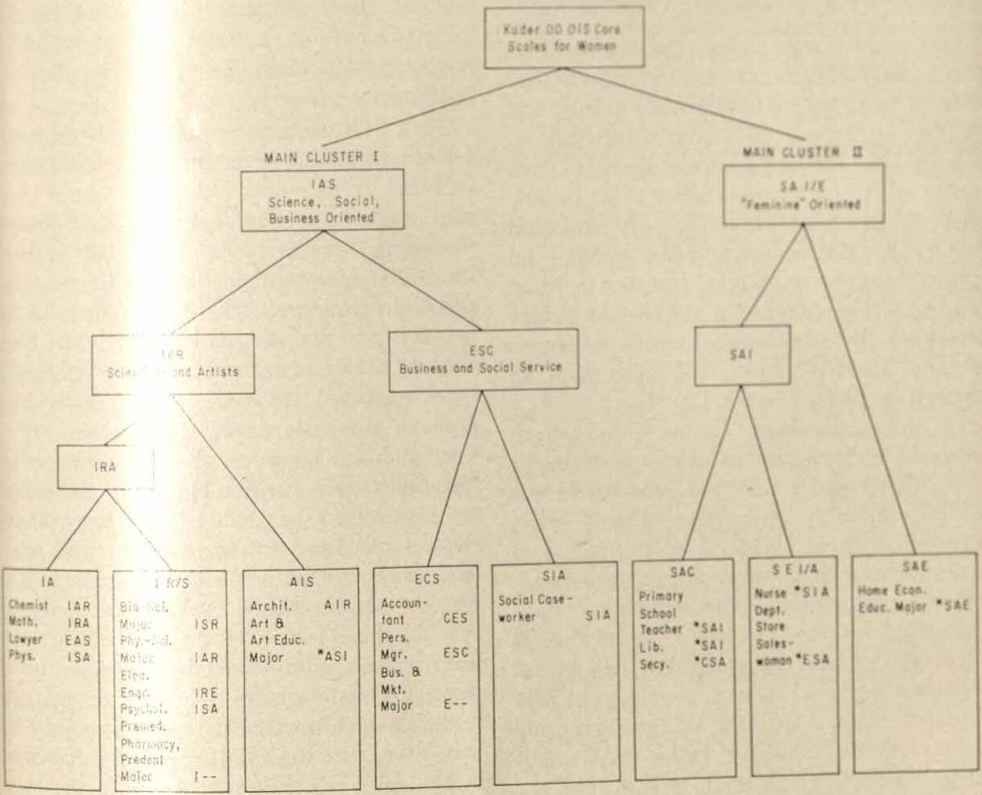


FIGURE 6. Main Clusters I and II for Kuder DD Occupational Interest Survey (OIS) core scales for women with abbreviated group summary codes and descriptive labels. (Asterisk denotes scales derived from female criterion groups.)

Kuder DD OIS Clusters for Women

The cluster analysis for the OIS core scales for women is shown in Figure 6. The two groups resulting from the first major division were an IAS cluster of 15 occupations and an SA I/E cluster of 6 occupations. The scales were divided into a heterogeneous group of science, social, and business occupations and a more homogeneous group of traditional "feminine" occupations.

Main Cluster 1—Science, Social, and Business Oriented

For the OIS core scales, the pattern of clustering in Main Cluster 1 for women was very similar to Main Cluster 2 for men, and the Holland codes seemed consistent with the pattern. Differences that did occur in the cluster patterns between men and women were confounded with the differences in the job titles included in the two forms of the OIS making a direct comparison difficult.

Division of Main Cluster 1 yielded a small cluster of business and social occupations, coded ESC, and a larger group of scientific and artistic occupations, coded IAR. The latter cluster divided predictably into one group of artistic jobs and two groups of scientific jobs. The Holland codes of the occupations within the subclusters were all closely related. One unexpected result was that lawyer (EAS) was grouped with chemist (IAR), mathematician (IRA), and physician (ISA). When the cluster of business and social occupations was divided, social caseworker (SIA) separated from accountant (CES), business and marketing major (E--), and personnel manager (ESC).

Main Cluster 2—"Feminine" Occupations

Two points concerning the occupations within Main Cluster 2 are important: (*a*) None of these appear in the men's profile, and all are traditional "feminine" occupations. (*b*) All of the scales were derived from female criterion groups (Kuder, 1971, p. 56). With the exception of art and art education major, all OIS women's scales normed on female groups were grouped together in Main Cluster 2. The remainder of the scales were developed by comparing women's responses to male criterion groups, and these were all grouped in Main Cluster 1.

On this basis, Main Cluster 2 seemed to represent a homogeneous group of women's occupational scales that was reflected in the SA I/E summary code. An analogous result occurred in Main Cluster 1 of the OIS core scales for men that contained primarily "masculine" occupations. The remaining subclustering for the women's scales seems somewhat unclear. In particular, two inconsistencies occurred where secretary (CSA) clustered with two SAI occupations, and home economics education major (SAE) split off by itself at a high level. In any case, the six occupations in the cluster seemed to be highly homogeneous relative to the other main cluster.

DISCUSSION

The results of this study provided evidence for the validity of Holland's occupational classification in three ways. First, occupations in the empirical clusters from the SVIB, MVII, and the OIS for both men and women possessed similar Holland codes. Second, the clusters suggested that a few broad groups of occupations exist that are internally consistent. Third, the hierarchical structure of the clusters followed the hexagonal ordering of the occupational categories.

Interpretation of the results of the analyses must be done cautiously. Individually and collectively, the instruments that were analyzed included relatively few occupations, so all interrelationships that occur in the world of work were not necessarily accounted for in these analyses. In addition, the clusters were largely instrument specific regarding the questionnaire items and scoring methods. Despite these precautions, a great deal of similarity was apparent among the results for the four instruments.

The SVIB contains scales for a larger number of occupations than the other instruments analyzed in the study. Other than the R category, each of Holland's types is adequately represented by the SVIB. At higher levels of the cluster hierarchy for the SVIB, classifications into groups occurred with respect to the first letter of the occupational code. In large clusters where two or more of Holland's types appeared the types were usually hexagonally contiguous and, thus related. Divisions at lower levels on the cluster hierarchy sometimes were predictable from the second or third letters of the Holland code. Many clusters consisted of occupations with Holland codes of the same three letters but in various permutations. This result suggests that all three letters in the Holland occupational code are important in describing occupational characteristics. That all occupations in the clusters were not perfectly homogeneous may have been the result of misclassifications in Holland's scheme, deficiencies in the measurement instrument, or the inability of the cluster-analytic method to adequately resolve differences in the data.

The cluster analysis of the MVII provided support for the notion that, within any one category of Holland's classification, occupations may be further subcategorized in a meaningful way by using second and third letters of the codes. The MVII scales are primarily Realistic occupations and, in general, represent that category of occupations.

When the MVII scales clustered, those that involved contact with people stayed together, and those that involved the manipulations and modification of materials stayed together. The former group of occupations required more social competencies and interests, while the latter required more investigative competencies and interests. The second letters of Holland codes reflected this difference in terms of the social and investigative components in the group summary codes. The results of the MVII analysis offered the clearest support of the utility of the second and third letters of the Holland code.

Results for clusters of the OIS core scales were parallel for men and women: each had one main cluster that appeared to be sex specific; the other main cluster was similar for both sexes. The sex-related cluster for men included primarily Realistic occupations. For women the sex-related cluster was mostly composed of Social occupations. In both cases these clusters were consistent with popular masculine–feminine stereotypes, and reflected the male–female roles that are found in our culture.

The main cluster that was similar for both sexes consisted of occupations that were science, social service, and business oriented. Further clustering and subclustering were much the same for the men and women and were consistent with results that were expected in terms of Holland's codes.

REFERENCES

BORGEN, F. H., & WEISS, D. J. Cluster analysis and counseling research. *Journal of Counseling Psychology*, 1971, 18, 583–591.

CAMPBELL, D. P. *Handbook for the Strong Vocational Interest Blank*. Stanford, Calif.: Stanford University Press, 1971.

CLARK, K. E., & CAMPBELL, D. P. *Manual: Minnesota Vocational Interest Inventory*. New York: Psychological Corporation, 1965.

HOLLAND, J. L. *Manual: Vocational Preference Inventory*. (Rev. ed.) Palo Alto, Calif.: Consulting Psychologists Press, 1970.

HOLLAND, J. L. *Making vocational choices: A theory of careers*. Englewood Cliffs, N.J.: Prentice-Hall, 1973.

HOLLAND, J. L., VIERNSTEIN, M. C., KUO, H., KARWEIT, N. L., & BLUM, Z. D. A psychological classification of occupations. *JSAS Catalog of Selected Documents in Psychology*, 1972, **2**, 84. (Ms. No. 184)

HOLLAND, J. L., WHITNEY, D. R., COLE, N. S., & RICHARDS, J. M., JR. *An empirical occupational classification derived from a theory of personality and intended for practice and research*. (ACT Research Report 29) Iowa City, Iowa: American College Testing Program, 1969.

KUDER, G. F. *General Manual: Kuder DD Occupational Interest Survey*. (Rev. ed.) Chicago: Science Research Associates, 1971.

MCQUITTY, L. L., & CLARK, J. A. Clusters from iterative, intercolumnar correlational analysis. *Educational and Psychological Measurement*, 1968, **28**, 211–238.

(Received May 10, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 354–357

CHECKING GROUPS OF LETTERS[1]

IAN E. GORDON[2] AND MOIRA AMOS

University of Exeter, England

In Experiment 1, 25 subjects searched through paired groups of letters to check for identity; nonidentical groups (10%) were to be marked. The response times increased with the number of letters per group (from 2 to 6) and with horizontal separation between the pairs of groups (1, 2, 4, and 8 centimeters). The optimum number of letters for efficient error checking appeared to be 4 or 5. In Experiment 2, 10 subjects searched through some of the Experiment 1 lists. Horizontal eye movements were recorded. Results indicated that: (*a*) the time spent looking at groups increased with the number of letters per group, (*b*) during error checking the second member of a pair of groups was looked at longer than the first, and (*c*) commonly, with small groups, more than one group was processed at a time.

Strings of digits and/or letters appear increasingly as labels. For instance, a typical example of the new British postal codes is EX3 OEU. The present report forms part of an investigation into the effects of various spatial arrangements of such codes on efficiency of search. In the present study interest is confined to all-letter groups and the situation investigated is error checking. Data on error-checking groups of digits have been published by Klemmer (1959) who found that the maximum rate of checking digits occurs when they are grouped into 3s or 4s.

The two experiments reported were designed (*a*) to find the optimum size of groups for letter checking, (*b*) to investigate the effects of increasing the spatial separation between horizontal paired groups, and (*c*) to examine subjects' scanning patterns by recording their eye movements during search.

EXPERIMENT 1

Method

Subjects. The 25 subjects were undergraduates at the University of Exeter. Their mean age was 19 years 7 months.

Material. Groups of random lower case letters typed in Remington elite were reproduced with good black/white contrast by Xerox. Twenty search lists were constructed, each consisting of 50 paired groups of random letters forming two columns 21.5 centimeters high mounted symmetrically about the vertical mid-line of a 20 × 33 centimeter white card. Thus each card contained two columns of digits. Letters were typed single-spaced. Pairs of letter groups were identical except for five pairs in every list that differed by a single letter. The vertical position of groups containing these wrong letters was randomized and varied from list to list. The position of the wrong letters within groups was also randomized.

The number of letters per group varied from two to six. Groups were printed using four different horizontal separations: 1, 2, 4, and 8 centimeters. These figures refer to the distance between the rightmost letter of the left group and the leftmost letter of the adjacent right group. With the typeface used the horizontal extent of a six-letter group was 1.3 centimeters.

Procedure. The subject worked through all 20 lists (5 Groupings × 4 Spacings) arranged in random order. His task was to work down a list comparing paired groups. A check mark was placed against any pair that did not match. Times per list and omissions were recorded. A 20-second rest was given between successive lists, and a 2-minute rest was given after the tenth list. The subjects were tested individually and sat with the search list placed on a horizontal surface in front of them. They started to search as soon as a list was uncovered, and they indicated when they had reached the end of the list. The instructions placed emphasis on speed and accuracy. Knowledge of results was not provided.

Results and Discussion

The results obtained in Experiment 1 are summarized in Figure 1 and discussed in detail below.

Group size. The overall mean search times across all the spacings for the two- to six-letter groups were: 34, 38, 44, 56, and 73 seconds, respectively. However, the aim of

---

[1] This research was supported by Grant B/SR 7092 from the Science Research Council.

[2] Requests for reprints should be sent to I. E. Gordon, Department of Psychology, The University, Exeter, England.

Experiment 1 was to discover optimum groupings, and so the search times/list were divided by the total number of letters/list to provide a more appropriate measure—letter-processing rate.

Figure 1 shows changes in the processing rate recorded in Experiment 1. It is clear that within the range of conditions studied, the fastest processing occurs when letters are grouped into fours or (with eight-centimeter spacing) fives. The influence of group size on rate is statistically significant ($F = 80$, $df = 4/96$, $p < .001$). This is in good agreement with Klemmer's demonstration that groupings of three and four are optimal for rapid checking of digits.

Letters and digits. After the completion of Experiment 1, and at the suggestion of an anonymous referee, one of the checklists used was reconstructed using random digits in place of letters. Apart from the change to digits, the new list was identical with the five-group, four-centimeter list previously described. Ten new undergraduate subjects, mean age 21 years, searched through this list on two occasions. The procedural details in this small study were identical with those previously described. The resulting overall mean search time was 55 seconds, which is in strikingly good agreement with the time for the comparable letter condition (56 seconds).

Spacing between groups. Reference to Figure 1 shows that, generally, the greater the spacing between groups the slower is the letter-processing rate. This is a relatively small effect, at least over the range of separations employed, as the difference between overall average processing rates at one-centimeter separation and eight-centimeter separation is only 1.2 letters/second. The spacing effect is statistically significant ($F = 35$, $df = 3/72$, $p < .001$) and is probably due to the extra time required to prepare and execute saccadic movements over greater distances.

Errors. All recorded errors were errors of omission. The average number of omissions across conditions was 13.4%. Errors and search times/list were related ($\rho = +.76$, $p < .01$), thus larger groups took longer to search and were searched less accurately.

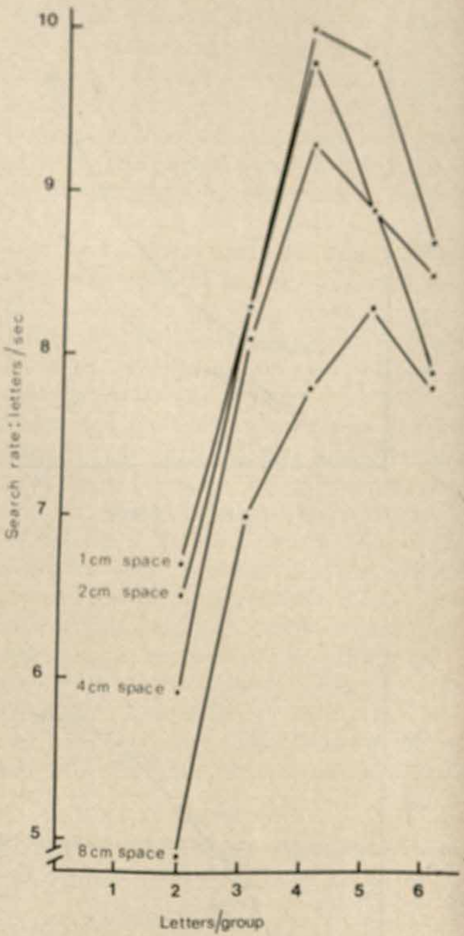


FIGURE 1. Search rate as a function of group size and group spacing in Experiment 1. Each point represents the average of 25 subjects.

EXPERIMENT 2

This experiment was designed after analysis of data from Experiment 1. Given the limited extent of foveal vision, it is obvious that when the spatial separation between two groups is increased beyond a certain distance, eye movements are required in the comparison of the groups. In Experiment 2 the widest separation (8 centimeters) used in Experiment 1 was again used so as to require the subject to make large, easily monitored saccades.

Experiment 2 was designed to find out (a) how the average duration of visual fixations

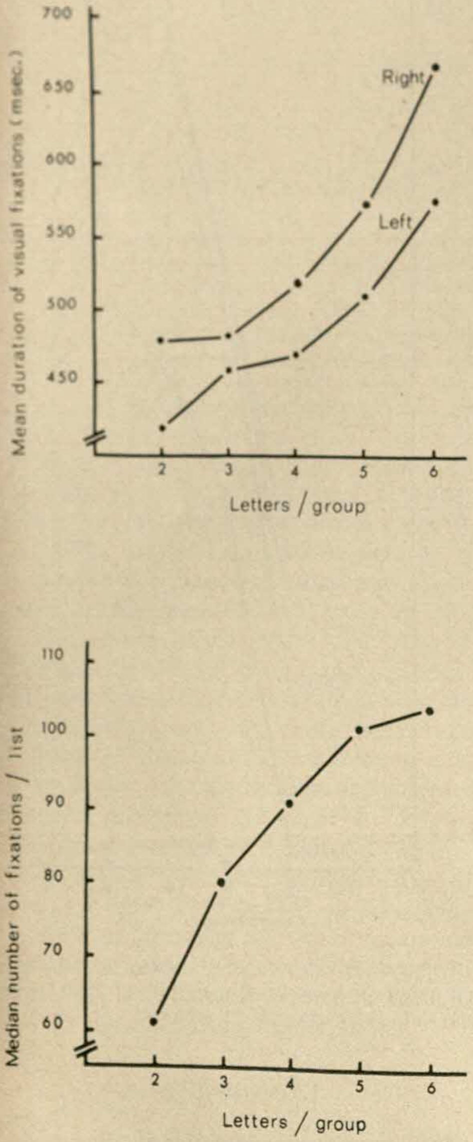


FIGURE 2. Eye movement responses to changes in group size in Experiment 2. Each point represents the average of 10 subjects.

varies with the size of a letter group, (*b*) which member of a pair of groups is fixated longer, the first or second, (*c*) whether any first/second differences in fixation time vary with group size, and (*d*) to what extent increases in overall search time with larger group sizes are due to the subject fixating longer or more frequently.

Method

Subjects. Ten undergraduates from the University of Exeter served as subjects. Their mean age was 21 years 4 months.

Material. Lists containing paired groups of two to six letters, all with eight-centimeter separation were used. They were identical with lists used in Experiment 1.

Procedure. The experimental details were very similar to those described in Experiment 1. Each subject was tested individually on all five lists presented in random order. The subject's head was supported but not constrained by a chin rest. The card bearing the search material was mounted on a table in front of the subject. The card was below the subject and inclined at 45 degrees. The distance from the centre of the card to the subject's eyes was 40 centimeters.

Instead of marking each error item (as in Experiment 1), the subject pressed a button that placed a mark on his eye movement record. He was given a 20-second rest between successive lists. The subject was specifically instructed not to turn back to reexamine earlier pairs in a list.

During search, horizontal eye movements were monitored by electrooculography. A silver electrode was placed at the external canthus of each eye. Eye movements were analyzed by direct measurement of the output wave form from a pen recorder fed by a high-gain direct current amplifier.

Results and Discussion

The eye movement data from Experiment 2 are unambiguous and are reported fairly briefly. The main horizontal eye movement patterns observed during error checking are summarized in Figure 2.

Preliminary trials with the investigators as subjects showed that, with a horizontal separation of eight centimeters between groups, there was never any difficulty in deciding from the eye-movement wave form which side of the display (left or right group) was being looked at. Electrooculography does not record very small microsaccades, and vertical eye movements were not recorded. Therefore the term "fixation" used below and in Figure 2 means only that the gaze was unambiguously directed to one or other side of the display and that no detectable horizontal eye movement was occurring.

Fixation time. As the number of letters/group was increased, the average time between large saccades increased. This effect is statistically significant ($F = 7.7$, $df = 4/36$, $p < .01$), and suggests that there is a sequential component in the processing of items during error checking.

Number of fixations/list. Increasing the number of letters/group produced an increase in the number of visual fixations made during a search down a list. This effect is best described in terms of median numbers of fixations per list, and these are illustrated in Figure 2. The result is statistically significant ($F = 11.9$, $df = 4/36$, $p < .01$). It is clear also that when the subject checked groups containing four letters or less he did not separately fixate every group. Indeed, with groups of only two letters the median number of fixations made during comparisons of 100 groups per list was only 61.

Left versus right groups. From Figure 2 it can be seen that in all conditions right groups are fixated longer than are left groups. This effect is statistically significant ($F = 7.3$, $df = 1/9$, $p < .05$), and the overall difference in mean fixation times to left and right groups is approximately 60 milliseconds. During error checking the first group (usually the left member of a pair) must be scanned and remembered and then a stored representation is compared with the scanned input from the second group. The process of comparison might happen in two ways: (*a*) Items from first and second groups might be compared sequentially and if this occurred during a fixation pause it would explain why right fixation times tend to be longer. (*b*) The subject might compare items simultaneously—thus whilst scanning groups of letters appears to be a sequential process, comparison after scanning might be a parallel process. This second conclusion is supported by the fact that the upper pair of curves in Figure 2 are roughly parallel—if comparison time were a direct function of the number of items compared the curves would be expected to diverge as group size increased.

The present results, together with those of Klemmer (1959), lead to two practical recommendations: (*a*) If digits or letters are to be grouped to facilitate rapid error checking, groupings of three to five items result in optimum performance; use of group sizes outside this range depresses performance. (*b*) Within the limits of a normal printed page, increasing the spacing between adjacent groups slows performance; this effect is likely to remain a small one, provided that groups which are to be checked are printed on the same page or are otherwise brought within about eight centimeters of each other. However, placing groups within one centimeter of each other facilitates search and checking.

REFERENCE

KLEMMER, E. T. Numerical error checking. *Journal of Applied Psychology*, 1959, **43**, 316–320.

(Received March 13, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 358–364

EFFECTIVENESS OF HIGHLIGHTING FOR RETENTION OF TEXT MATERIAL[1]

ROBERT L. FOWLER,[2] AND ANNE S. BARKER
University of South Florida

The effectiveness of "highlighting" selected text material for improving retention was examined in 4 groups of 19 college students each. Although there were no differences among the groups in total score on a retention test one week after reading with different highlighting techniques or no highlighting, a detailed analysis implied that highlighting improves retention of selected text material and that active highlighting is superior to passive reading of highlighted material. Highlighting seemed most effective when the reader had maximum faith that the highlighter could discriminate between important material and trivia. Traditional underlining also was found effective as an emphasis technique.

Textbooks often employ typographical cueing (capital letters, italics, and colored type) for parts of the lesson assumed to have greater importance or difficulty for the student. Some research has indicated that such typographical cueing results in an improvement in retention of the emphasized material when compared with material printed in standard homogeneous typeface (Hershberger & Terry, 1965; Klare, Mabry, & Gustafson, 1955). In spite of these findings, the widespread use of typographical cueing has not been apparent in educational materials, either because of cost factors or the reluctance of authors and publishers of textbooks to designate some of the material as unimportant and deserving of only superficial reading.

In the absence of typographical cueing, students are often faced with the task of deciding which portions of text material are important enough to be isolated for intensive study or memorization. A recent survey by the authors revealed that of 200 randomly selected used textbooks for sale in a local college bookstore, some 92% contained significant applications of "emphasizing techniques" that had apparently been performed by the previous readers on the text material. Among the more popular of these techniques were the uses of underlining, asterisks, parentheses, checkmarks, and brackets, usually in a color or brightness level that contrasted with the text print. The most recent isolation method in widespread use in the textbooks examined is known as highlighting, or accenting. Highlighting changes the background color of a line of black-on-white print through the stroke of a special marking pen of a bright but relatively unsaturated hue such as yellow, blue, or red. The highlighting procedure is normally performed with each student serving as his own judge as to which and how much material should be isolated for study emphasis. However, a number of cases have been observed in which a low-achieving student has attempted to improve his test performance by "studying" from a textbook previously highlighted by a high-achieving student.

Two experiments were carried out to examine the effectiveness of various highlighting techniques. The purpose of Experiment 1 was to determine (*a*) whether or not highlighting is effective as an aid to the retention of instruction material, (*b*) whether or not active participation in the highlighting task enhances the presumed beneficial effect, (*c*) the degree of agreement among students reading educational material for the first time regarding what portions of text should be emphasized, and (*d*) the relationship between the proportion of material chosen for emphasis and retention.

Experiment 2 compared the facilitating effects of two specific "emphasis" techniques as aids to retention. Several previous studies have shown the effectiveness of underlining in reading (Cashen & Leicht, 1970; Idstein & Jenkins, 1972), but highlighting has apparently not been compared with underlining in reading. Since the highlighter pen is manufactured specifically for use as an emphasizer and sells for approximately twice the price of a regular ball point pen, this question is of some interest.

---

[1] This study was facilitated by a University of South Florida Research Council Release Time Grant to the first author.

[2] Requests for reprints should be sent to Robert L. Fowler, University of South Florida, 830 First Street South, St. Petersburg, Florida 33701.

EXPERIMENT 1

Method

Subjects

The subjects were 78 undergraduate male and female students enrolled in general psychology courses. Participation was voluntary, but students were promised extra points toward the course grade based on retention test performance. Since the text material consisted of course-related outside readings relevant to the students' interests, intrinsic motivation was assumed to be relatively high as well.

Materials

The text readings were reprints from the *Scientific American* (Heron, 1957) and *Science* (Milgram, 1970) dealing with the pathology of boredom and the social perils of city life. These proved to be inherently interesting, yet basically unfamiliar to the students. The length of the articles together was about 8,000 words and was appropriate for a one-hour initial study period.

The highlighter pen used was a relatively unsaturated yellow (Major Accent #2500) which when stroked produced a colored line of approximately one-eighth-inch width superimposed on each emphasized line of text. This instrument was chosen because it is the most widely used highlighter in the immediate area and because the yellow (approximately 570 nm) is near the point of maximum spectral sensitivity for photopic vision (Forgus, 1966).

The retention test consisted of 54, 4-alternative multiple-choice items. The items were constructed in order to produce the appropriate balance between highlighted and nonhighlighted material for the group for which this was the relevant independent variable. Eighteen test items were based on highlighted statements, and the remaining 36 on nonhighlighted statements. It was presumed that this test provides an opportunity for highlighting to produce an effect on total score; that is, 33% of the items were drawn from only 20% of the material. Yet, ample opportunity existed for the student-highlighted groups (described below) to improve their test performance by selecting for highlighting part or all of the material from which the remaining 67% of the items was drawn. Based on the administrations reported in Experiments 1 and 2, the test yielded a Kuder-Richardson reliability of .78.

Design

Each subject was randomly assigned to one of four groups. One group actively highlighted while studying (Group AHL). Each subject in this group was allowed to highlight as much of the text as he deemed appropriate. In a second group, each subject was "yoked" to a member of Group AHL in the sense that he studied from the copy of the material previously highlighted by his yoke mate, although these subjects were not told who had done the highlighting. This was then, the passive highlighting group (Group PHL). When a student assigned to Group AHL failed to highlight any of the text material, he was permitted to complete the test but was not assigned a yoked control, and his score was not included in the data analysis. One subject was excluded from the analysis on this basis and another Group AHL subject was excluded before yoking because of his bizarre and unresponsive highlighting technique.

A third group studied copies of the material previously highlighted by the experimenter (Group EHL). The ratio of highlighted to nonhighlighted material was 1:4 and was based on a survey of textbooks used by honor students locally. The fourth group was a control group that studied the same material without any highlighting (Group NHL). Other than for the appropriate amount and type of highlighting, all four groups were treated identically.

Procedure

All study, review, and testing sessions were conducted in individual sound-insulated cubicles. Although the assignment of subjects to experimental conditions was as random as possible, obviously, the subjects in Group PHL were run after their yoke mates.

The subject was seated alone at a table in one of the cubicles. On the table was placed the appropriate reading material with a blank cover sheet. The experimenter then read standard instructions which informed the subject that he would be given one hour to study the articles at this time and would return in exactly one week for a 10-minute review followed by a test. Groups PHL, EHL, and NHL were admonished not to mark on their copies of the articles, while Group AHL was shown how to use the highlighter pen for "emphasizing material you may think is particularly important."

At the end of the one-hour study period, the experimenter retrieved the reprints and asked the subject to refrain from discussing the material or procedure with anyone and to return in one week. The title and reference location of the articles had been removed from all copies to reduce the possibility that any subject could obtain and read the articles during the intervening week. One week later, the subject was seated in the same cubicle as

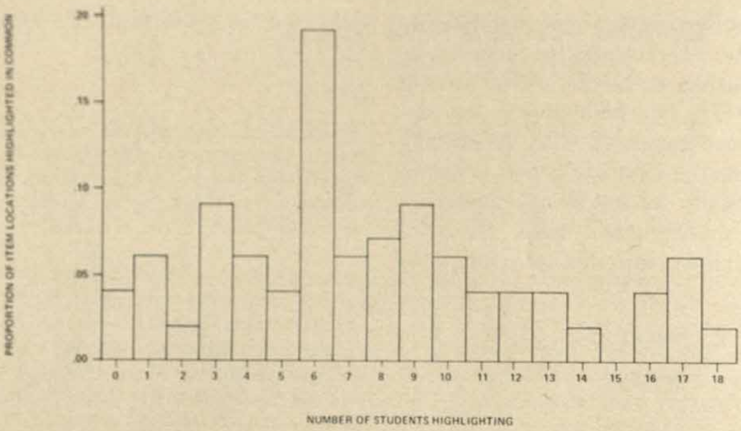


FIGURE 1. Proportion of material highlighted in common by a given number of students.

before and given the same copy of the articles for review. No highlighting or other marking was permitted during this period. After 10 minutes of review, the subject was given the 54-item objective examination with instructions to answer all items in a maximum time of 50 minutes.

Results

A completely randomized analysis of variance failed to reveal reliable differences in retention test scores among any of the four experimental groups on this measure ($F < 1$, $df = 3/72$). However, a comparison of Groups EHL and NHL on just those 18 items drawn from material prehighlighted for Group EHL showed the former ($\bar{X} = 13.3$) to be significantly superior to the latter ($\bar{X} = 11.8$, $t(36) = 2.04$, $p < .05$). Thus, Group EHL enjoyed an advantage on these selected items, but it was not sufficient to significantly increase their total score over that of the control Group NHL. An examination of the variability of the scores within groups called into question the power of the experimental design employed to detect differences which might exist among treatment conditions. Therefore it was decided to perform a more detailed analysis on an item-by-item basis.

Since the subjects in Group AHL differed widely with respect to the proportion of material they chose to highlight (range = 4.2% – 32.1%, $\bar{X} = 16.2\%$) and consequently, varied in the amount of test item material they happened to emphasize for test review, it was necessary to examine the individual Group AHL booklets to determine the actual effects of active versus passive highlighting. For example, some test items came from text material highlighted by almost all Group AHL subjects, while other items were drawn from passages highlighted by no one; the mean proportion of Group AHL subjects highlighting the answer to a given test item was .41. The distribution of proportion of item material highlighted in common by a given number of Group AHL subjects is given in Figure 1.

This item-by-item analysis of each Group AHL booklet made it possible to compare the performance of Groups AHL and PHL separately for highlighted and nonhighlighted item material. By employing a randomized-blocks factorial analysis of variance (Kirk, 1968) with items as the blocking factor (considering only those 52 items from material marked by at least one Group AHL subject), the probability of passing an item for both groups combined, given that the answer to it had been highlighted ($\bar{X} = .72$), was significantly greater than if it had not been highlighted ($\bar{X} = .62$, $F = 17.17$, $df = 1/153$, $p < .001$), although the mean proportion of items passed did not differ significantly for the two groups ($F = 2.33$, $df = 1/153$, $p > .10$). Interestingly, when Tukey's HSD test for pairwise comparisons (Kirk, 1968) was employed subsequently, an item was passed more often within the Group AHL when it had been highlighted ($\bar{X} = .76$) than when

TABLE 1

PROPORTION OF ITEMS PASSED BY EACH GROUP UNDER EACH CONDITION

Group	*n*	Item condition					
		Highlighted	Items[a] involved	Not highlighted	Items[a] involved	Adjacent	Items[a] involved
AHL	19	.76	417	.62	609	—	—
PHL	19	.69	417	.61	609	—	—
EHL	19	.74	342	.68	684	.73	190
NHL	19	(.66)[b]	(342)[b]	.69	1,026	(.76)[b]	(190)[b]

Note. AHL = actively highlighting; PHL = passive highlighting; EHL = experimenter highlighted; NHL = no highlighting.

[a] No. items involved summed over all subjects in group.

[b] Items highlighted for Group EHL.

it had not ($\bar{X} = .62$, $p < .01$), but the same was not true within Group PHL. Additionally, the proportion passing an item, given that it had been actively highlighted, was significantly greater than the proportion passing all items in Group NHL ($\bar{X} = .69$, $z = 2.59$, $p < .01$), when Fisher's test of significance of a difference between proportions was employed (Guilford, 1965). The same comparison involving Groups PHL and NHL showed the proportions passing to be not significantly different. The means for each group under each condition are given in Table 1.

In an ancillary analysis, which was performed to test the effects of highlighting on retention of adjacent material (spread-of-emphasis effect), the 10 items taken from material adjacent to material prehighlighted for Group EHL were examined separately. It was determined that for Group EHL, adjacent items were passed significantly more often than were items which were neither adjacent nor highlighted ($t = 3.44$, $df = 18$, $p < .001$). However, these adjacent items could have been, as a group, less difficult than the remaining items, since Group NHL also passed them significantly more often ($t = 2.37$, $df = 18$, $p < .05$). The question of whether or not the superior retention of emphasized material occurs at the expense of the remaining text material cannot be addressed directly from the results of this study. However, the proportion passing nonhighlighted items was significantly less in both Group AHL ($z = 2.66$, $p < .01$) and Group PHL ($z = 3.30$, $p < .01$) when each was compared to the total item performance in control Group NHL using Fisher's test of significance of a difference between proportions (Guilford, 1965). The rank order of means in Table 1 also suggests that isolation may lead to better retention of the emphasized portions as well as concomitantly poorer retention of the remaining portions. This issue is further complicated by the observed superior retention of items from material adjacent to highlighted portions.

The relationship between the proportion of material highlighted by members of Group AHL and their total score on the test was not statistically significant ($r = -.287$, $df = 17$, $.20 > p > .10$), nor did the degree of this relationship differ for Groups AHL and PHL. However, the correlation between the proportion of Group AHL subjects highlighting a given item and the probability of their passing the item was reliably greater than zero ($r = .295$, $df = 52$, $p < .05$), although the same relationship for Group PHL was not ($r = .239$, $df = 52$, $p > .05$). As a further control for the possibility that the answers to the easier items tend to be highlighted, the relationship between the probability of a Group AHL subject highlighting the answer to an item and the proportion of Group NHL subjects passing the same item was found to be insignificantly greater than zero ($r = .082$, $df = 52$, $p > .90$).

EXPERIMENT 2

Method

Subjects

The subjects were 24 volunteers from a graduate course in principles of learning.

TABLE 2

Counterbalanced Orders of the Experimental Condititions for the First Three Matched Items

Order	Item number		
	13	14	43
A	underline	highlight	control
B	underline	control	highlight
C	control	underline	highlight
D	highlight	underline	control
E	highlight	control	underline
F	control	highlight	underline

Materials and Design

The same journal articles and retention test used in Experiment 1 were again employed. However, the emphasis markings on the booklets were applied differently in order to employ a within-subjects design comparing the effectiveness of red underlining, yellow highlighting, and no-emphasis control conditions. To accomplish this, the item analysis of the reading retention test administered in Experiment 1 was examined to identify multiple-choice items that had positive discrimination indexes and moderate difficulty indexes (.35 to .85). From these selections it was possible to match items by threes according to similarity of difficulty level. A total of 12 such triplets were available with each member of a triplet being assigned to one of the three experimental conditions for a particular subject in the following manner: (*a*) The text material comprising the answer to one of the items in a triplet was emphasized by highlighting; (*b*) that answering another was underlined; and (*c*) the answer to the third was left unemphasized. Since there are six possible permutations of the three experimental conditions across the three items of a triplet, each permutation was randomly assigned to four subjects' booklets in order to completely counterbalance this condition. Table 2 shows the arrangement of the experimental treatments for each of the six groups (orders) of four subjects for the first triplet of similar-difficulty items. The same procedure was followed for each of the remaining 11 triplets. On each booklet, additional emphasis was applied until a total of 10% of the material was underlined and 10% was highlighted. This was done to create a more realistic situation where not all emphasized material contained answers to test items.

Procedure

During the regularly scheduled Tuesday night meeting of the course, students were asked to remain after class if they wished to volunteer to participate in a "learning" experiment. Booklets containing an equal number of the six different orders of technique application were distributed randomly to the 24 students who remained. A cover sheet on each booklet contained instructions concerning the procedure to be followed. The subjects were again allowed 60 minutes to read the material originally and were told to study as they would for an exam but not to make marks on the booklet.

They were also informed that, on the following Tuesday night, they would be given the same copy of the articles to review for 10 minutes prior to a multiple-choice test based on the material. All subjects studied the material and took the test in the same room at the same time.

Results

Although the retention test contained 54 items, only the 36 items comprising the matched triplets were examined as dependent measures of performance. Thus, each subject's score could vary from 0 to 12 under each experimental condition. These data were analyzed as a randomized-blocks analysis of variance with each subject being a block (Kirk, 1968) and showed highly reliable differences in performance among the emphasis techniques ($F = 5.79$, $df = 2/46$, $p < .01$). A posteriori comparisons between each pair of means were performed with Tukey's HSD test (Kirk, 1968) and revealed the mean performance for underlining (8.61) to be significantly greater ($p < .05$) than for the nonemphasized control condition (7.13). The performance under the highlighted condition (8.04) did not differ significantly from that under either of the other two conditions. The improvement in performance due to underlining was approximately 21% which again corroborates the previous findings regarding the magnitude of this effect.

DISCUSSION

The question of whether or not highlighting, as an emphasizing technique, increases the retention of text material must be answered with a qualified "yes." In Experiment 1, the test items based on material highlighted were passed significantly more often than were the remaining items in every comparison performed. Yet the total test score means did not differ significantly among any of the four treatment groups. In the same study, the improvement in performance by highlighting over the appropriate nonhighlighted control condition ranged from about 12% to [illegible]%, compared to about 18% in a similar study employing underlining by Cashen and Leicht (1970); thus, allowing for the possibility that increased retention of emphasized material may *adversely* affect the retention of the remaining material, a result consistent with many studies dealing with the von Restorff effect in verbal learning reviewed by Wallace (1965), a liberal estimate of the overall increase in test scores due to emphasizing would be on the order of 6%–7%, which is considerably less than both the observed standard deviation of total test scores and the error of measurement due to test unreliability. This result could account for the failure of some investigators to find positive results in total test score measures when the test contains items from both emphasized and nonemphasized text (Idstein & Jenkins, 1972).

Perhaps the most interesting finding was the superior performance of Group AHL relative to Group PHL on those items based on material actively highlighted by the former and passively read by the latter. Since these groups did not differ in performance on nonhighlighted items, the implication is that active participation in the emphasizing process is necessary for optimum results. It may also be that students are more likely to mark material that is easier to retain, or that contains information with which they agree. Moreover, the superior retention of emphasized statements by Group AHL may simply reflect the fact that active highlighting of a text statement guarantees that it has been read, whereas passive highlighting does not. Thus, exposure rather than isolation may be the controlling factor.

The choice of material for active highlighting by the students did reflect a relatively high degree of discrimination between statements suitable for objective examination and those serving merely as connecting discourse, that is, the average of 16% total material marked contained over 40% of the retention-test-item answers. Nevertheless, the large within-group variance on this factor plus the fact that the greater portion of retention test material was not highlighted may have contributed to the relatively poorer performance on the highlighted items in the passive condition.

One of the controversies regarding the effectiveness of isolation as a retention aid in prose material concerns the role of the reader of the material. Crouse and Idstein (1972) maintain that the encoding (by underlining) of text material is effective if done *for* the reader but not if done *by* the reader since he must guess what to encode while being uninformed about the contents of a subsequent (retention) test. This interpretation is consistent with the positive results obtained in experiments where material is preemphasized (Cashen & Leicht, 1970; Crouse & Idstein, 1972) and with the weak or negative effects resulting when readers do their own emphasizing (Idstein & Jenkins, 1972; Stordahl & Christensen, 1956). The results of Experiment 1 support this view in that the four groups did not differ in terms of total test score performance, whereas Group EHL did outperform control Group NHL on those items prehighlighted for the former. The degree of enhanced retention of isolated material by passive readers may ultimately depend on the amount of faith they have in the judgment of the person who did the highlighting. Unlike the von Restorff effect obtained in serial lists where retention of all elements in the list is required, the theory of emphasis employed in the present study assumes the differential importance of the various text segments and that some portions should be retained to a greater extent, even at the expense of the remaining portions. It follows that a student studying from a prehighlighted text must believe that the more important statements have been emphasized on his copy or he may ignore, rather than concentrate on, the highlighted portions. In fact, a bias could develop in the early stages of studying an assignment when the student notices a discrepancy between his judgment of what is important and what he observes on the prehighlighted text.

The data of Experiment 2 suggest that the more expensive and special-purpose highlighter pen is no more effective as an emphasizer than are more traditional techniques such as a ballpoint pen containing ink of a different color than the text material print.

This observed relative ineffectiveness of the highlighter pen is consistent with the weak effect found under the passive conditions in Experiment 1. Perhaps a different outcome would be obtained if the subjects were allowed to select their own emphasis technique and then actively apply it while reading the material.

Thus, the findings confirm that highlighting can produce some improved retention and that traditional underlining of portions of text material may be as good as use of the more expensive highlighter pens. The evidence suggests further that readers of instructional material containing little or no typographical cueing may benefit from highlighting done previously by others (especially if done by students seen as high achievers). However, students retain more if they do the highlighting themselves.

REFERENCES

CASHEN, V. M., & LEICHT, R L. Role of the isolation effect in a formal educational setting. *Journal of Educational Psychology,* 1970, **61**, 484–486.

CROUSE, J. H., & IDSTEIN, P. Effects of encoding cues on prose learning. *Journal of Educational Psychology,* 1972. **63**, 309–313.

FORGUS, H. *Perception*. New York: McGraw-Hill, 1966.

GUILFORD, J. P. *Fundamental statistics in psychology and education*. New York: McGraw-Hill, 1965.

HERON, W. The psychopathology of boredom. *Scientific American,* 1957, **196**(30), 52–56.

HERSHBERGER, W. A., & TERRY, D. F. Typographical cuing in conventional and programed texts. *Journal of Applied Psychology,* 1965, 49, 55-60.

IDSTEIN, P., & JENKINS, J. R. Underlining versus repetitive reading. *Journal of Educational Research,* 1972, **65**, 321–323.

KIRK, R. E. *Experimental design: Procedures for the behavioral sciences*. Belmont, Calif.: Brooks/Cole, 1968.

KLARE, G. R., MABRY, J. E., & GUSTAFSON, L. M. The relationship of patterning (underlining) to immediate retention and to acceptability of technical material. *Journal of Applied Psychology,* 1955, **39**, 40–42.

MILGRAM, S. The experience of living in cities. *Science,* 1970, **167**, 1461–1468.

STORDAHL, K. E., & CHRISTENSEN, C. M. The effect of study techniques on comprehension and retention. *Journal of Educational Research,* 1956, **49**, 61–70.

WALLACE, W. P. Review of the historical, empirical, and theoretical status of the von Restorff phenomenon. *Psychological Bulletin,* 1965, 63, 410–424.

(Received April 9, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 365–368

A FACTOR ANALYSIS OF AIR PASSENGER REACTIONS TO SKYJACKING AND AIRPORT SECURITY MEASURES AS RELATED TO PERSONAL CHARACTERISTICS AND ALTERNATIVES TO FLYING[1]

MICHAEL R. COOPER[2]

GTE Laboratories, Inc., Waltham, Massachusetts

CHARLES E. BOLTWOOD AND ROBERT J. WHERRY, SR.

Ohio State University

A 28-item questionnaire administered to 127 airline passengers at Columbus International Airport, Ohio, was factor analyzed and rotated to orthogonal hierarchical structure. Results indicated that passengers did respond differentially to items concerning their reactions to hypothetical skyjacking situations and airport security measures when items were couched in terms of willingness to fly as contrasted to probability of actually taking a given flight. Subgroups were differentiated by one or more of the following variables: age, sex, enjoyment of flying, distance to destination, and the extent of their dislike for buses. Other than buses, preferred modes of travel appeared relatively unrelated to favorableness or unfavorableness of passenger reactions.

Two recent studies (Boltwood, Cooper, Fein, & Washburn, 1972; Cooper, Fein, Washburn, & Boltwood, 1971) have shown that different passenger groups reacted differentially to the elaborate security measures designed to combat skyjacking. The present study attempts to investigate the extent to which factors accounting for these differential reactions could be explained by means of personal variables or passenger preferences for various modes of transportation.

METHOD

A 28-item questionnaire was constructed to measure the following variables:

- A. Willingness to fly after seeing or hearing about
  (1) metal detection devices, (2) armed security guards, (3) frisking, (4) general security level, (5) a recent skyjacking, (6) a passenger death due to skyjacking, (7) past threat of skyjacking.
- B. Probability of still taking a trip after seeing or hearing about
  (8) metal detection devices, (9) armed security guards, (10) frisking, (11) general security level, (12) a recent skyjacking, (13) a passenger death due to skyjacking, (14) past threat of skyjacking.
- C. Personal Characteristics
  (15) sex, (16) age, (17) enjoyment of flying, (18) distance to destination.
- D. Preferred Mode of Travel for Short Trips
  (19) Bus, (20) Car, (21) Boat, (22) Train, (23) Plane.
- E. Preferred Mode of Travel for Long Trips
  (24) Bus, (25) Car, (26) Boat, (27) Train, (28) Plane.

The questionnaires were randomly administered at the Columbus International Airport, Ohio, over a one-week period to 127 passengers within the general waiting area. Only two persons refused to be interviewed. The sample comprised a variety of flights, both short and long, for passengers of both sexes and a wide age span.

---

[1] This study was conducted while the first author was a research associate in the Organizational Research Service, Department of Psychology, Ohio State University. The views expressed in this article are those of the authors and do not necessarily reflect the views of GTE Laboratories, Inc., or Ohio State University.

[2] The authors wish to thank Victoria Fein Bernstein for her invaluable assistance during the preliminary phases of the study, Thomas M. Ostrom for his comments on the construction of the questionnaire, and the late Milton D. Hakel, S. Rains Wallace, and Michael T. Wood for their constructive criticisms of earlier drafts of this report. Also, the cooperation of the Federal Aviation Administration and Daniel Ginty of Columbus International Airport is gratefully acknowledged.

Requests for reprints should be sent to Michael R. Cooper, Behavioral Science Research, GTE Laboratories, Inc., 40 Sylvan Road, Waltham, Massachusetts 02154.

TABLE 1

Willingness, Probability, and Personal Characteristic Factors

Factor	A	1	2	3	4
Willingness to Fly					
1. Metal detectors	*30*	*41*	−06	−18	16
2. Armed guards	*46*	*62*	07	−02	−07
3. Frisking	*49*	*48*	−01	*25*	−10
4. General security	*34*	*46*	*29*	−18	−08
5. Skyjacking, recent	*40*	10	−02	*40*	00
6. Skyjacking, severe	*46*	16	06	*36*	08
7. Skyjacking, past threat	16	*−25*	00	*27*	*19*
Probability of Flying					
8. Metal detectors	*25*	−18	−04	11	*38*
9. Armed guards	*43*	*25*	*33*	01	12
10. Frisking	*48*	*22*	*37*	14	−02
11. General security	*36*	06	*71*	−07	−14
12. Skyjacking, recent	*41*	−15	*42*	*43*	00
13. Skyjacking, severe	*49*	−11	*55*	*23*	*22*
14. Skyjacking, past threat	05	*−37*	10	−10	*50*
Personal Characteristics					
15. Sex	*−35*	*−21*	07	*−21*	−16
16. Age	*35*	−05	−07	−05	*28*
17. Enjoyment of flying	−03	14	03	14	*−22*
18. Distance to destination	−07	*−24*	*18*	*−24*	−13
Mode Preference					
Short Trip					
19. Bus	*−23*	03	07	03	*−46*
23. Plane	06	02	−05	−08	*21*

Note. Variables with no loading greater than .11 have been omitted.

The variables were intercorrelated and factor analyzed (principal factor plus minimum residual to clean up communalities). The eight factors extracted (highest residual equal to .09 and lowest eigenvalue equal to .30) were subjected to a varimax rotation followed by a hierarchical rotation (Wherry, 1959). The hierarchical rotation resulted in the introduction of three subgeneral factors.

The reader should bear in mind that the findings of this study apply only to airline passengers and not to the general public or even to travelers in general.

Results and Discussion

Willingness, Probability, and Personal Characteristic Factors

Significant factor loadings on the questionnaire items related to willingness to fly and probability of flying were restricted to the first and largest subgeneral factor with four of the primary factors clustered under it. The loadings for these factors appear in Table 1.

Subgeneral Factor A—General Favorable Reaction to Situations. This factor is composed primarily of older males who disliked buses for short trips and who had little reaction of any kind to the past threat of skyjacking. They were either so favorably impressed with the various security measures designed for their protection or so little frightened by reported skyjackings (either recent or severe) that they still reported a high willingness to fly and a high probability of taking their flights. The subfactors under this Subgeneral A originally all showed this same pattern to some degree; however, that overlap has been removed by the hierarchical rotation.

Factor 1—Relatively High Willingness to Fly. This subgroup of males traveling to nearby destinations, who had reacted more negatively to the past threat of skyjacking were apparently more *willing* to fly when metal detectors, armed guards, frisking, and a general high-security level were in evidence. Their nearby destinations may have involved smaller planes less apt to be skyjacked to a foreign country.

Factor 2—Relatively High Probability of Taking a Flight. This subgroup of persons about to embark on relatively longer flights, who had responded neither negatively nor positively to the past threat of skyjacking were responding more highly with respect to actual probability of taking their flights than in terms of willingness to fly. Their probability of flying was increased by all signs of protection, with the exception of metal detectors, and especially by the general level of security. They had not responded much to the past threat of skyjacking and were not affected by hypothetical reports of recent or severe skyjackings.

Factor 3—Relatively Favorable Reaction Despite Skyjacking. This subgroup of men embarking on relatively short trips indicated relatively high willingness and probability of flying with respect to both the past threat of skyjacking and the two hypothetical skyjackings presented. Their willingness to fly was increased more by seeing passengers frisked than by other signs of security precautions.

Factor 4—Relatively Favorable Response Despite Past Skyjackings. This subgroup of relatively older passengers who did not like to fly but nevertheless preferred planes to buses for short trips responded particularly favorably despite past skyjackings which had taken place. Their probability of flying was increased when metal detection devices were used.

Thus, three factors indicate older men are less apt to be intimidated by either past or present (hypothetical) skyjackings or by signs of increased security activity. The four subfactors indicate that age, sex, enjoyment of flying, distance to destination, and dislike for buses may cause certain subgroups to respond differentially to the various threats or precautions with respect to their willingness or determination to actually fly.

Mode of Transportation Factors

The remaining two subgeneral factors, each associated with two primary factors, are mostly related to varying preferences for modes of travel as related to distance of destination. One of the subgenerals is minimally related to willingness to fly in view of past or severe skyjackings with one of its primaries related to flying enjoyment. Loadings are given in Table 2.

Subgeneral Factor B—Preference for Planes over Boats or Cars for Long Trips. This group of passengers, who still showed slightly higher willingness to fly in the face of past or severe skyjacking reports, preferred planes over either boats or cars for long trips (over 300 miles). The two subordinate factors merely show increased antagonism toward either boats or cars but not toward both.

Factor 5—Preference for Planes over Cars. This subgroup indicates further preference for planes over cars. Their long trips may have been intracontinental where boats would not have been a feasible alternative.

Factor 6—Preference for Planes over Boats. This subgroup possibly represents persons who are planning trips across oceans to foreign countries where boats would have been the only alternative.

Subgeneral Factor C—Boats over Trains for Long Trips; Planes or Cars over Trains for Short Trips. This subgroup is primarily marked by a dislike for trains, especially for long trips. The two group factors represent smaller subgroups of individuals.

Factor 7—Planes over Cars for Short Trips. This subgroup, when boats are not available, expresses a strong preference for planes over cars for short trips only. Trains may not have been available for the short trips in question.

TABLE 2

CHOICES OF TRAVEL MODE FOR SHORT AND LONG TRIP FACTORS

Factor	B	5	6	C	7	8
Willingness to Fly						
6. Skyjacking, severe	*23*	14	−04	03	−08	−09
7. Skyjacking, past threat	*21*	14	12	−04	−04	10
Personal Characteristic						
17. Enjoyment of flying	00	−01	−16	−01	−12	*43*
Mode Preference						
Short Trip						
20. Car	00	−05	03	*−20*	*−83*	*49*
22. Train	10	−15	13	*−23*	−02	*−71*
23. Plane	10	−01	−03	*38*	*84*	01
Long Trip						
25. Car	*−50*	*−57*	−04	−06	−04	15
26. Boat	*−22*	−10	*−34*	*42*	05	10
27. Train	−09	−02	−00	*−31*	−08	*−57*
28. Plane	*57*	*59*	*44*	−01	15	*38*

Note. Variables with no loading greater than .16 have been omitted.

Factor 8—Cars over Trains for Short Trips; Planes over Trains for Long Trips. This subgroup, who incidentally like to fly very much, when boats are not available and in view of their strong dislike for trains, choose cars over trains for short trips and planes over trains for long trips. The choice of cars rather than planes as the alternative to trains for short trips may result from the fact that actual flying time for these trips may be less than the time spent in getting to and from the airport, which may be nearly as great as simply driving.

All of the aforementioned mode of transportation factors indicate a strong preference for planes over other modes of transportation. Indeed, 89% of the sample did rank planes highest among all possible modes of travel for long distances. The nature of the trip, short versus long or continental versus transoceanic, seems to determine the choices available for travel and/or the various advantages and disadvantages of the alternative modes of travel.

Overall, the results indicate the following: (*a*) Older males who were opposed to using buses tend to be more willing and more likely to fly despite actual or hypothetical skyjackings or signs of airport security. (*b*) Males who had been negatively affected by past skyjackings and were making trips to nearby destinations were more willing to fly when there was a generally high-security level and when frisking, use of metal detectors, and especially when armed guards were in evidence. (*c*) Persons who were little affected, one way or the other, by previous skyjackings were more likely to actually take their flights when they observed a high level of security precautions, including frisking and the presence of armed guards. (*d*) Men who were embarking on relatively short trips showed more willingness and were actually more likely to fly despite past actual skyjackings or hypothetical skyjackings. They seemed to be particularly assured by the evidence of frisking. (*e*) Older passengers who did not like to fly but were strongly opposed to traveling by bus seemed to be least disturbed by past skyjackings and most reassured by the use of metal detectors. (*f*) Planes were preferred over cars and/or boats for long trips by persons whose willingness to fly was not shaken by previous, even severe skyjacking incidents. (*g*) Some members of the sample preferred boats over trains for long trips. (*h*) Certain members of the sample preferred planes over cars for short trips. (*i*) Other members, who were especially fond of flying and preferred planes over trains for long trips, nevertheless preferred cars over trains for short trips.

REFERENCES

Boltwood, C. E., Cooper, M. R., Fein, V. E., & Washburn, P. V. Skyjacking, airline security, and passenger reactions: Toward a complex model for prediction. *American Psychologist,* 1972, 27, 539–545.

Cooper, M. R., Fein, V. E., Washburn, P. V., & Boltwood, C. E. Passenger attitudes toward airline security and threat of skyjacking. Paper presented at the meeting of the American Psychological Association, Washington, D.C., September 1971.

Wherry, R. J. Hierarchical factor solutions without rotation. *Psychometrika,* 1959, 24, 45–51.

(Received May 3, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 369–370

SHORT NOTES

REPLICATION OF WHITE-COLLAR–BLUE-COLLAR DIFFERENCES IN SOURCES OF SATISFACTION AND DISSATISFACTION

THOMAS C. HARRIS AND EDWIN A. LOCKE[1]

University of Maryland

Small samples of white-collar and blue-collar employees were asked to describe sources of satisfaction and dissatisfaction on the job. These critical incidents were classified using the Schneider-Locke event-agent classification system. It was found, consistent with previous findings, that white-collar employees were more likely to derive satisfaction and dissatisfaction from "motivator" events (especially achievement/failure), and blue-collar employees from "hygienes" (especially money).

Previous studies have indicated that white-collar employees more often get pleasure from intrinsic or motivator factors and blue-collar employees from extrinsic or hygiene factors (Herzberg, Mausner, Peterson, & Capwell, 1957; Locke, 1973). The present study attempted to replicate these results using the Herzberg critical incident method and the Schneider & Locke (1971) event-agent classification system.

METHOD

The subjects were 31 white-collar and 14 blue-collar employees from three organizations. Fourteen white-collar subjects were professionals, while the remainder were clerical. Fourteen white-collar employees were male. All of the blue-collar subjects were male; 13 were bookbinders, while 1 was a carpenter. The mean tenure of both groups was 4–5 years. The mean age of the blue-collar group was higher (43 years versus 30 years). Internal analyses indicated, however, that neither the age nor the sex differences between the samples could account for the differences reported below.

Each subject was asked anonymously to describe one satisfying and one dissatisfying incident and to classify the main event and agent that characterized each incident. The Schneider and Locke event categories were: task activity, amount of work, smoothness of work, achievement/failure, promotion, responsibility, recognition, money, interpersonal relations, and working conditions (for complete descriptions, see Locke, 1973). The first seven categories were classified as "motivators" and the last three as "hygienes" (Herzberg, Mausner, & Snyderman, 1959). Since Grigaliunas and Herzberg (1971) indicated that subjects cannot always be trusted to categorize accurately, the incidents were recoded independently by the two authors into event categories. (Agent recategorization was not done, due to lack of sufficient detail in the stories.) There was 80% agreement on the individual categories, but 92% agreement as to whether the incident was a motivator or a hygiene event. Disagreements were resolved by discussion. The investigators' coding agreed with the subjects' coding only 48% of the time based on individual categories, but 87% of the time based on the motivator/hygiene classification.

RESULTS

The 2 × 2 frequency analyses of motivator/hygiene by satisfaction/dissatisfaction for both white-collar and blue-collar employees and for both subjects' classifications and investigators' classifications yielded no significant chi-squares. This indicates that, for both employee groups and both methods of classification, the same events were seen as leading to both satisfaction and dissatisfaction. Thus for the remaining event analyses, the satisfying and dissatisfying incidents were combined.

Table 1 shows the frequency of motivator and hygiene events for the white-collar and blue-collar samples. Both the results based on the subjects' own classifications and those based on the investigators' classifications are shown. Despite the subject–experimenter disagreements concerning the coding of individual incidents, it is evident that the overall frequencies are similar

---

[1] The authors wish to thank William Jeffries and William Roach who aided in the collection of the data for this study.

Requests for reprints should be sent to Edwin A. Locke, College of Business and Public Administration, University of Maryland, College Park, Maryland 20742.

TABLE 1

MOTIVATORS VERSUS HYGIENES AS SOURCES OF SATISFACTION AND DISSATISFACTION FOR WHITE-COLLAR AND BLUE-COLLAR EMPLOYEES

Causes of satisfaction and dissatisfaction	Employee	
	White-collar	Blue-collar
Motivator event	52^a	15^a
	52^b	17^b
Hygiene event	10^a	13^a
	10^b	11^b
χ^2	9.31^{a**}	
	5.78^{b*}	

[a] Authors' coding.
[b] Subjects' coding.
$^* p < .05$.
$^{**} p < .01$.

for both methods of coding. The chi-squares in both cases were significant at $p < .05$ or better, indicating that white-collar employees are more likely to report getting pleasure and displeasure from motivator events and less likely to get these experiences from hygiene events than are blue-collar employees.

This overall difference was caused almost exclusively by the more frequent mention of achievement/failure events among white-collar employees and the more frequent mention of monetary events among the blue-collar workers. These data are shown in Table 2.

With respect to the data for agents, white-collar employees had a significant tendency to take credit for satisfying events and to blame others for dissatisfying events ($\chi^2 = 10.15$, $p < .01$). This result did not replicate with the blue-collar employees. There was no overall difference (combining satisfying and dissatisfying incidents) between the blue-collar and white-collar samples with respect to attributing events to the self versus to others.

TABLE 2

FREQUENCY OF MENTION OF ACHIEVEMENT/FAILURE AND MONEY AS SOURCES OF SATISFACTION AND DISSATISFACTION

Causes of satisfaction and dissatisfaction	Employee	
	White-collar	Blue-collar
Achievement/failure	26^a	1^a
	12^b	1^b
Money	2^a	7^a
	3^b	5^b
χ^2	21.43^{a*}	
	7.29^{b*}	

[a] Author's coding.
[b] Subject's coding.
$^* p < .01$.

DISCUSSION

These data were congruent with previous findings regarding white-collar–blue-collar differences. The failure to replicate Herzberg's two-factor theory when the event–agent classification system is used also agreed with previous findings (Locke, 1973; Schneider & Locke, 1971). Finally, in one respect the results replicated the findings of Grigaliunas and Herzberg (1971): that subjects and trained coders may disagree substantially in their coding of incidents. On the other hand, the disagreements in the present study were predominantly *within* the motivator or hygiene categories rather than across these categories. The result (Table 1) was substantial overall agreement when the individual categories were grouped according to the motivator/hygiene dichotomy. In contrast, there was substantial disagreement *across* motivator/hygiene categories in the Grigaliunas and Herzberg study. This may have been due, in part, to the logical inconsistency in the Herzberg classification system which confuses agents with events (see Schneider & Locke, 1971, pp. 442–443).

REFERENCES

GRIGALIUNAS, B. S., & HERZBERG, F. Relevancy in the test of motivator-hygiene theory. *Journal of Applied Psychology*, 1971, **55**, 73–79.

HERZBERG, F., MAUSNER, B., PETERSON, R. O., & CAPWELL, D. F. *Job attitudes: Review of research and opinion*. Pa.: Psychological Service of Pittsburgh, 1957.

HERZBERG, F., MAUSNER, B., & SNYDERMAN, B. *The motivation to work*. New York: Wiley, 1959.

LOCKE, E. A. Satisfiers and dissatisfiers among white-collar and blue-collar employees. *Journal of Applied Psychology*, 1973, **58**, 67–76.

SCHNEIDER, J., & LOCKE, E. A. A critique of Herzberg's incident classification system and a suggested revision. *Organizational Behavior and Human Performance*, 1971, **6**, 441–457.

(Received April 16, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 371-373

CORRELATES OF JOB SATISFACTION IN NAVAL ENVIRONMENTS[1]

BLAIR W. McDONALD AND E. K. ERIC GUNDERSON[2]

Navy Medical Neuropsychiatric Research Unit, San Diego, California

Demographic, military status, job-related and health-related variables were correlated with a job satisfaction measure for 5,851 Navy enlisted men. Health-related variables were most highly correlated with satisfaction, followed by type of job and seniority. A multiple regression equation with six variables (symptom index, type of job, length of service, specific ship, number of men supervised, and worked closely with others) yielded a cross-validity coefficient of .49. The results indicated some important correlates of job satisfaction in the naval environment and suggested hypotheses for further investigation.

The determinants and consequences of job satisfaction in military settings have received relatively little attention except in the area of career retention. In fact, very few studies have been conducted of job satisfaction in enlisted personnel since the large-scale Army studies of World War II (Stouffer, Suchman, DeVinney, Star, & Williams, 1949).

The purpose of the present study was to determine some of the important correlates of job satisfaction as reported by Navy enlisted men in operational settings. Further, it was felt that an investigation of this type might provide a basis for more detailed investigations aimed at improvements in the level of satisfaction within the Navy.

METHOD

Subjects

Subjects were 5,851 enlisted Navy personnel aboard six ships (three cruisers, two aircraft carriers, and one battleship). Five of the ships were deployed to Southeast Asia, and the sixth ship participated in training exercises in the Mediterranean area. The duration of deployment for the six ships ranged from 6 to 8 months. The subjects were tested enroute to their overseas stations under the supervision of experienced research teams. Approximately 80%–90% of the enlisted personnel aboard the cruisers and the battleship participated in the study, while random samples were drawn from the carrier crews. Approximately one half of the subjects had two years or less of military service and were under 21 years of age. Slightly more than 60% of the subjects had completed high school and about 30% were married.

Procedure

Personal history, military status, and job satisfaction information were obtained from a single questionnaire. All questionnaires were checked for completeness and appropriateness of response, and about 3% were discarded because they were incomplete or invalid.

The measure of job satisfaction used for the major study was composed of three items. The key item, that is, the one given greatest weight in scoring, was: "How satisfied are you with your present job?" The response alternatives for this item were: "Very much," "Slightly satisfied," "Slightly dissatisfied," and "Very dissatisfied"; these were scored 4, 3, 2, 1, respectively. The second most important item was: "Do your present duties employ your abilities in the best possible way?" The response alternatives were: "Not at all," "Partly," and "Very much," scored 1–3. The third item—"Are you often bored?"—had only "Yes" (scored 1) or "No" (scored 2) as response alternatives. The response values for the three items were summed to obtain the total job satisfaction score. The internal consistency reliability estimate, based on average item intercorrelations and corrected by the

---

[1] Report Number 73-23, supported by the Bureau of Medicine and Surgery, Department of the Navy, under Research Work Unit MF51.524.002-5015DX5F. Opinions expressed are those of the authors and are not to be construed as necessarily reflecting the official view or endorsement of the Department of the Navy.

[2] The assistance of CDR R. H. Rahe, MC, USN, Biochemical Correlates Division, Navy Medical Neuropsychiatric Research Unit, and R. T. Rubin, formerly of the Research Unit, in data collection and interpretation of the results is gratefully acknowledged. W. M. Pugh and Jeanne M. Erickson performed the statistical analyses for the study.

Requests for reprints should be sent to E. K. Eric Gunderson, Navy Medical Neuropsychiatric Research Unit, Department of the Navy, San Diego, California 92152.

Spearman-Brown formula, was considered adequate for a brief scale ($r_{xx} = .70$).

Seventeen variables reflecting demographic, military status, and job-related information all were derived from the questionnaire, including scores on a brief symptom index, the Health Opinion Survey (Gunderson, Arthur, & Wilkins, 1968). An index of family socioeconomic status was derived from the fathers' education and occupation. The number of illnesses (dispensary visits) during the overseas deployment period was determined from ships' medical records. Variables concerned with duty or job type and specific ship were created by means of a linearization procedure, that is, ordering response categories in terms of mean criterion (job satisfaction) scores.

RESULTS

The Navy job satisfaction scale correlated .61 with the Smith Job Descriptive Index (JDI) scale in a large sample of Navy men aboard ship ($n = 1,021$). This relationship was of the same magnitude as that between the JDI and Porter's Need Satisfaction scale (Imparato, 1972). In view of the fact that the JDI was not specifically designed for military populations and settings, the relationship with the Navy job satisfaction measure provides evidence of substantial agreement with respect to the general concept being measured. The relationships between JDI and the Navy satisfaction measures are described in detail elsewhere.

Relationships between the 17 variables reflecting personal, military, job, and health status, and job satisfaction were tested for significance by means of Pearson product-moment correlations. Of the 17 correlations, 16 were significant ($p < .05$). Only racial or ethnic group, that is, Caucasian versus non-Caucasian (primarily Negro), was not significantly correlated with job satisfaction.

Satisfaction scores were positively correlated with age, length of service, and pay grade ($r = .24$, .23, and .26, respectively, $p < .01$). Satisfaction varied directly with the number of close co-workers ($r = .20$ and .15, respectively, $p < .01$). Satisfied men tended to be married ($r = .06$, $p < .05$) and to have dependents ($r = .15$, $p < .01$). Satisfaction was negatively correlated with the subject's level of education and the socioeconomic status of parents ($r = -.09$ and $-.06$, respectively, $p < .05$).

With regard to the type of work performed, higher satisfaction was reported by men who were working in clerical, ordnance, or cook/steward jobs ($r = .12$, .11, and .09, respectively, $p < .01$). Subjects in temporary duty (for example, compartment cleaner) or deck force assignments reported the lowest job satisfaction ($r = -.21$ and $-.16$, respectively, $p < .01$).

In order to evaluate further the factors predictive of job satisfaction, a multiple regression analysis in a cross-validation design was used. The total sample was stratified on the basis of age and education and a sample of 2,474 subjects was drawn for the validation portion of the analysis. A regression equation consisting of six variables was developed against job satisfaction as the criterion. The variables which contributed significantly to the multiple correlation ($R = .49$) were the Health Opinion Survey score, type of job, length of service, the specific ship that the subject was serving aboard, the number of men supervised, and working closely with others. When this regression equation was applied to the cross-validation sample of 2,458 subjects, a cross-validity correlation coefficient of .49 was obtained.

DISCUSSION

The cross-validity coefficient obtained with six predictor variables was substantial and demonstrated that satisfaction was predictable to a moderate degree from the variables of this study. The variable that correlated most highly with job satisfaction was the Health Opinion Survey scale ($r = -.32$), a brief symptom index that has been shown to discriminate psychiatric patients from healthy controls in the naval service (Gunderson, Arthur, & Wilkins, 1968). This result, coupled with the fact that actual illness (dispensary visits) correlated negatively with satisfaction ($r = -.12$, $p < .01$), strongly suggested that health status is significantly involved in job attitudes. The schedule of recent experience, which has been shown to be related to health in a series of studies (Rahe, 1972), also correlated significantly with satisfaction ($r = -.18$, $p < .01$). The schedule of recent experience measures recent changes in the individual's personal, social, and occupational life that are reported to be psychologically stressful. It appears plausible that the "wear and tear" of life stresses outside of the job situation affect job satisfaction to some extent just as such stresses affect health.

The next most important set of variables related to satisfaction reflected naval experience (age and length of service) and level of responsibility (pay grade). The number of men supervised was clearly related to pay grade and experience. Working closely with others was correlated with but not entirely dependent on seniority or number of men supervised. This variable contributed uniquely to the prediction of job satisfaction in the regression equation, thus indicating that opportunity for close interpersonal relations was important for satisfaction among some of the sailors.

Aside from seniority and level of responsibility, the individual's particular occupational specialty was of some importance in determining level of satisfaction. The jobs generally considered most routine, arduous, dirty, or hazardous (temporary duty, deck force, and engineering and hull) were least satisfying.

Differences in job satisfaction from ship to ship were small but significant ($r = .12$, $p < .01$), and because these differences were independent of other variables, the specific ship variable contributed to prediction of overall satisfaction. Large differences in illness rates were previously noted for the same ships (McDonald & Gunderson, 1973) and, together with the present findings, suggested that organizational factors such as habitability conditions, level of proficiency, skills required, supervisory ratio, and other facets of organizational climate may be of importance in determining both health and satisfaction aboard ship.

REFERENCES

GUNDERSON, E. K., ARTHUR, R. J., & WILKINS, W. L. A mental health survey instrument: The Health Opinion Survey. *Military Medicine*, 1968, **133**, 306–311.

IMPARATO, N. Relationship between Porter's Need Satisfaction Questionnaire and the Job Descriptive Index. *Journal of Applied Psychology*, 1972, **56**, 397–405.

MCDONALD, B. W., & GUNDERSON, E. K. *Organizational factors and health status*. (Report 73-34) San Diego, Calif.: Navy Medical Neuropsychiatric Research Unit, 1973.

RAHE, R. H. Subjects' recent life changes and their near-future illness susceptibility. In Z. J. Lipowski (Ed.), *Advances in psychosomatic medicine*. Basel, Switzerland: Karger, 1972.

STOUFFER, S. A., SUCHMAN, E. A., DEVINNEY, L. C., STAR, S. A., & WILLIAMS, R. M. *The American soldier: Adjustment during Army life*. Vol. 1. New Jersey: Princeton University Press, 1949.

(Received March 26, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 374–376

RELATIONSHIPS BETWEEN PERFORMANCE AND SATISFACTION UNDER CONTINGENT AND NONCONTINGENT REWARD SYSTEMS

GERALD A. KESSELMAN,[1] MICHAEL T. WOOD,[2] AND EILEEN L. HAGEN

Ohio State University

Relationships between performance and satisfaction with specific job facets were examined for two samples of female telephone company workers. In the contingent system, pay was based on piece rates and promotions were tied to performance; in the noncontingent system, these rewards were based on seniority. Performance levels were similar under the two systems. However, in the performance-contingent sample, performance was related to satisfaction with the work itself, pay, and promotions. In the seniority sample, performance was related to satisfaction with interpersonal factors, pay, and work.

A major concern of organizational psychologists has been to determine the relationship between job behavior and job attitudes. Comprehensive reviews of early studies relating job satisfaction and job performance (Vroom, 1964) indicated that if satisfaction is related to performance, the degree of association is rather weak. Recent studies have considered conditions under which these two variables should be related (Schwab & Cummings, 1970).

The hypothesis that performance causes satisfaction suggests that reward contingencies mediate relationships between these two variables. In a laboratory experiment, Cherrington, Reitz, and Scott (1971) found that when rewards were contingent on performance, performance and satisfaction were positively related. When rewards were delivered randomly to high performers and low performers, performance and satisfaction were unrelated. Negative correlations were obtained when low performers were rewarded and high performers were not. Other experiments have shown higher levels of performance by subjects who were told that their earnings were contingent on the effectiveness of their performance (Atkinson, 1958).

In the present field study two hypotheses were tested: (*a*) Performance is greater under a reward system that explicitly rewards performance with pay than that under a system which administers pay noncontingent on performance. (*b*) Performance is positively related to satisfaction with pay, promotions, and work under a reward system that explicitly rewards performance with pay and promotional opportunities. This hypothesis assumes that performance is related to satisfaction with job facets specifically incorporated in the reward contingency system. Thus, increased performance in the work itself should associate with work satisfaction, since work effectiveness determines rewards. Satisfaction with pay and promotions should relate to performance since those specific rewards are contingent on performance.

METHOD

Subjects

The subjects were 37 draftswomen and 39 toll operators in a midwestern telephone company. Their median age was 29.5 years and median service was 5.4 years.[3] Draftswomen performed final artistic work on engineering blueprints and figures for equipment, cables, etc. While the two groups performed similar jobs with respect to work hours, training, and evaluation methods, their reward systems differed. The toll operators constituted the "seniority" group, where pay and promotions depended on length of service with the company, as defined by the union-company contract. Draftswomen in the "performance" group operated on the same contract combined with a piece-rate bonus system, where rewards were directly tied to job performance. The bonus system was based on individual weekly productivity above the standards determined by average worker productivity and task difficulty. The bonus was administered every three months in addition to regular hourly pay.

---

[1] This article is based on a portion of a thesis by the first author submitted to the Ohio State University Graduate School in partial fulfillment of the requirements for the master's degree.

Requests for reprints should be addressed to Gerald A. Kesselman, Department of Psychology, Ohio State University, 404C West 17th Avenue, Columbus, Ohio 43210.

[2] Now at Battelle, Human Affairs Research Centers, Seattle.

[3] The two samples were similar in terms of age, length of service, and salary. The education of draftswomen was greater than that of operators; however, the education variable was not related to any of the measures of interest in this study.

Measures

Satisfaction with pay, the work itself, promotions, co-workers, and supervision were measured by the Job Descriptive Index (Smith, Kendall, & Hulin, 1969) which was included in a larger questionnaire on job attitudes. The performance measure consisted of supervisors' ratings of individual performance that are completed semiannually and retained in company files. These ratings were on a 5-point Likert-type scale anchored by "poor job performance" and "superior performance."[4]

RESULTS

Hypothesis 1 was not supported. Levels of performance under the performance ($M = 2.97$) and seniority ($M = 3.13$) reward systems were similar.

Prior to testing Hypothesis 2, it was necessary to determine whether the subjects were actually being sampled from two distinct reward systems. Included in the questionnaire were items assessing the perceived contingency of pay and promotions on job performance. In the performance sample, 94.6% responded "Yes," while 2.7% responded "No," and 2.7% "?" to the question: "Is promotion based on job performance?" Among seniority subjects, 7.7% responded "Yes," 59% "No," and 33.3% "?." Subjects also responded to the following statement on a 5-point scale ranging from "strongly disagree" to "strongly agree": "In my job, the amount of pay I receive is primarily based on my job performance." Means were 4.11 and 2.03, respectively, for the performance and seniority subgroups ($t = 9.73$, $df = 74$, $p < .01$). Thus, pay and promotions were viewed as significantly more related to job performance in the performance sample than in the seniority sample.

The results in Table 1 generally supported Hypothesis 2. Although all correlations in the performance sample were positive, the only significant relations were between performance and satisfaction with work, pay, and promotions. In the seniority sample performance and satisfaction were also, generally, strongly interrelated. However, in this situation, performance was more strongly related to satisfaction with interpersonal components of the job, supervisor, and co-workers than was the case for the performance-contingency sample. Promotion satisfaction did not correlate with performance as strongly as in the performance sample. Contrary to expectation, pay and work satisfaction were highly related to superiors' performance ratings.

TABLE 1

CORRELATIONS BETWEEN PERFORMANCE AND SATISFACTION BY SAMPLE

Sample	Satisfaction variable				
	Work	Co-workers	Pay	Promotion	Supervisor
Total	.44**	.34**	.46**	.18	.33**
Performance	.58**	.27	.39*	.44**	.21
Seniority	.53**	.44**	.51**	.26	.49**

* $p < .05$.
** $p < .01$.

DISCUSSION

The results of the present study did not support the hypothesis that performance is greater under a reward system that explicitly rewards performance with pay than that under a system which administers pay noncontingent on performance. It must be reiterated, however, that this study was correlational, and reward contingencies were not experimentally manipulated. Although the two groups were similar in some organizational and personal respects, any differences between their behaviors could be attributed to differences on variables besides the reward systems, which were not measured. Moreover, the partial overlap of reward systems may account, in part, for the lack of observed performance differences; that is, in the performance group, only rewards beyond seniority-determined rates were contingent on productivity, and the reinforcement delay was substantial.

Studies which have been cited in support of the hypothesis have been conducted in the laboratory where reinforcement schedules and time intervals between performance and subsequent reward were under the control of the experimenter (Cherrington et al., 1971; Yukl, Wexley, & Seymore, 1972). As operant theory suggests, an important parameter of the performance-reward contingency is the time interval between performance and reward (Kimble, 1961, Ch. 6).

---

[4] For the performance sample only, measures of unit output were available. This productivity measure was highly related ($r = .97$) to supervisors' performance ratings.

With greater intervals, effects of the performance–reward contingency on future performance will decrease. In an organization, pay and promotion may be contingent on performance, but usually this does not occur until some time after the performance. Further field studies with varying reward schedules are needed to confirm the moderating effects of the time variable.

The performance–reward contingency mediated performance–satisfaction relationships. These results suggest further research on relationships between performance and specific facets of job satisfaction under organizational conditions where different rewards depend on performance. For example, if performance were rewarded by some change in supervisory practice, performance might relate to satisfaction with supervision. Research might also examine properties of reward systems that would explain performance–satisfaction relations when the reward contingency is seniority. In some situations where rewards are tied to service, workers may perceive performance as a path to continued service and rewards. Additional variables, such as individual reward values and equity, must be taken into account to predict organizational behavior and attitudes in behavior-contingent-reward systems.

REFERENCES

ATKINSON, J. W. Towards experimental analysis of human motivation in terms of motives, expectancies, and incentives. In J. W. Atkinson (Ed.), *Motives in fantasy, action, and society*. Princeton, N.J.: Van Nostrand, 1958.

CHERRINGTON, D. J., REITZ, H. J., & SCOTT, W. E., JR. Effects of contingent and noncontingent reward on the relationship between satisfaction and task performance. *Journal of Applied Psychology*, 1971, **55**, 531–536.

KIMBLE, G. A. *Hilgard and Marquis' conditioning and learning*. New York: Appleton-Century-Crofts, 1961.

SCHWAB, D. P., & CUMMINGS, L. L. Employee performance and satisfaction with work roles: A review and interpretation of theory. *Industrial Relations*, 1970, **9**, 408–430.

SMITH, P. C., KENDALL, L. M., & HULIN, C. L. *The measurement of satisfaction in work and retirement*. Chicago: Rand McNally, 1969.

VROOM, V. H. *Work and motivation*. New York: Wiley, 1964.

YUKL, G., WEXLEY, K. N., & SEYMORE, J. D. Effectiveness of pay incentives under variable ratio and continuous reinforcement schedules. *Journal of Applied Psychology*, 1972, **56**, 19–23.

(Received April 18, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 377–379

RELATIONSHIP OF EARLY WORK CHALLENGE TO JOB PERFORMANCE, PROFESSIONAL CONTRIBUTIONS, AND COMPETENCE OF ENGINEERS[1]

H. G. KAUFMAN[2]

Polytechnic Institute of Brooklyn, New York

This study investigated the relationship of the work challenge experienced by 85 engineers at the beginning of their career to their subsequent job performance, professional contributions, and competence. Engineers whose initial job experience involved challenging technical work tended to contribute to knowledge relatively early and to maintain professional competence and good job performance during their careers. For those who possessed high technical ability, early challenge further reinforced professional contributions and competence later in their career.

There is accumulating evidence that being assigned to challenging work during the first job experience has a favorable effect on performance and attitudes in succeeding years. One of the few longitudinal career development studies was carried out among college graduates hired as management trainees by the American Telephone and Telegraph Company (Berlew & Hall, 1966). The study theorized that at least two factors contribute to the relationship between early work challenge and subsequent performance, namely: (*a*) the organization initially assigns the most capable individuals to the most demanding jobs; and (*b*) a first job that is demanding leads to better performance and success in subsequent years. It is likely that these two factors, both independently and in interaction, affect not only subsequent job performance but also professional competence and contributions to knowledge in one's field. The following hypotheses were investigated in the present study: (*a*) challenging work at the beginning of the career encourages good job performance as well as professional competence and contributions to knowledge later in the career; and (*b*) this early work challenge interacts with initial individual ability to affect job performance, professional contributions, and competence beyond the independent effects of ability and challenge.

METHOD

The sample included 85 engineers, from three technologically based organizations, who received their baccalaureate degrees between 1953 and 1955. All had approximately contemporaneous work experience for a period of 14 years. The degree of technical challenge was measured at the end of two years of job experience by estimates of time allocated to as well as importance rankings of challenging technical work activities, using a form developed by the Educational Testing Service (Hemphill, 1963). These estimates were collected from both engineers and their managers and resulted in a total of four measures of early technical work challenge. A convergent-discriminant validity analysis demonstrated that all four estimates measured the same construct (Kaufman, 1973). In addition, test measures of engineering ability as well as interest in theory and ideas were obtained from the engineers shortly after their employment began (Kaufman, 1972). Those engineers who left their organizations did not differ significantly in ability and interest from those who remained.

Criterion data were collected for three time periods designated as early (first 8 years of experience), middle (9–11 years), and recent (12–14 years). The middle period was selected as being just at or beyond the "half-life" of an engineering degree, which has been estimated at 10 years or less (Roney, 1966). Management ratings of each engineer's job performance were collected for each time period, primarily from personnel records. Engineers contribute to new knowledge via patents and publications (LeBold, Perrucci, & Howland, 1966), and consequently, the number of publications and patent disclosures was used to measure professional contributions

---

[1] This research was supported by Grant 91-34-69-23 from the Manpower Administration, U.S. Department of Labor, under the authority of Title I of the Manpower Development and Training Act of 1962, as amended.

[2] Requests for reprints should be sent to H. G. Kaufman, Department of Management, Polytechnic Institute of Brooklyn, 333 Jay Street, Brooklyn, New York 11201.

TABLE 1

RELATIONSHIPS BETWEEN EARLY WORK CHALLENGE AND CRITERIA

Criteria		Early work challenge		Work Challenge×Ability
		Mean r	Mean % increase in R^2 after accounting for individual differences	Mean % increase in R^2 after accounting for individual differences and challenge
Performance				
	Y_1	.21*	5.7*	2.1*
($n=85$)	Y_2	.19*	2.5*	4.3*
	Y_3	.21*	3.3*	1.6
Publications				
	Y_1	.27*	2.5*	1.2*
($n=85$)	Y_2	.15*	1.6*	.0
	Y_3	.11*	.3	5.2*
Patents				
	Y_1	.10	.0	.0
($n=85$)	Y_2	.21*	1.1	.0
	Y_3	.11*	.0	1.8*
Competence				
($n=24$)	Y_3	.32*	6.2*	7.0*
Diversity				
($n=24$)	Y_3	.34*	3.6*	6.4*

Note. The data presented are the means of the results obtained using four different measures of challenge. Y_1 = early period; Y_2 = middle period; Y_3 = recent period.

* At least one relationship significant at the $p \leq .05$ level.

during each career stage. In addition, an inventory of technical specialities, completed by the engineers in one of the organizations ($n = 24$), provided highly correlated measures of current technical diversity and competence in sophisticated professional knowledge (Kaufman, 1972). Competence, diversity, and patent disclosures were all significantly correlated with the number of publications during the recent period. With the exception of these relationships and the correlations between earlier publications and recent competence, the other criteria were not significantly correlated with each other during or across time periods.

The statistical analysis involved the obtaining of product-moment correlations between the measures of work challenge and the criteria for each of the three time periods. Multiple regression techniques were used to test for the increase in criteria predictability contributed by early work challenge as well as by the interaction between this challenge and individual ability after accounting for the effects of ability and interest (Cohen, 1968). The latter procedure was carried out since ability and interest predicted the assignment of engineers to challenging work (Hemphill, 1963).

RESULTS

A summary of the results appears in Table 1. Since the four utilization criteria are measures of the same construct, the means for correlation and percent increase in variance were used to summarize the results. The product-moment correlations between early work challenge measures and criteria in later periods were all positive, with many being significant, thus tending to support the first hypothesis.

Early work challenge consistently added to the prediction of subsequent job performance during both early and later periods. The interaction between initial ability and work challenge significantly contributed to the prediction of job performance during the early and middle periods, thereby providing some support for the second hypothesis.

The prediction of publications was significant with a clear tendency for the relationship between early work challenge and publications to be strongest during the initial period and to diminish progressively during the later periods. However, the prediction of patents was strongest during the middle or "half-life" period.

Providing further support for the second hypothesis was the finding that the interaction of individual ability and work challenge significantly contributed to the prediction of the number of publications and, to a lesser degree, the number of patent disclosures during the most recent period. In addition, both technical diversity and professional competence during the most recent period were significantly predicted by early work challenge treated both independently and in interaction with individual ability, thus providing additional support for both hypotheses.

DISCUSSION

The finding that early work challenge consistently predicted job performance, even in later stages of an engineer's career, may be attributed to the stimulation provided by the early challenge toward good initial performance. This very likely resulted in more challenging work assignments generating a continuous process of reinforcement. Such a speculation tends to be supported by evidence from another longitudinal study of engineers, indicating that relationships between organizational factors and performance remain consistent over time (Farris, 1969).

Unlike performance, professional contributions to knowledge may only be an immediate consequence of early challenging work with the time lag in disclosing patents appearing slightly longer than publishing articles. However, for those engineers initially possessing high technical ability, early challenging work may help to stimulate professional development later in the career. The contribution of early work experience to the independent criteria of job performance and professional development indicates that the effects of work challenge may be global and not specific to job or profession, whereas ability and interest may be more specific in their effects (Kaufman, 1972).

Although the emphasis on such techniques as job enrichment in recent years has focused mainly on nonprofessional employees, it is the professionals in the organization who are likely to be most affected by the intrinsic motivation of their work. Hence, organizations can initiate change in the job design of their professionals in order to enhance the development of the individual in both job and career (cf. Kaufman, 1974).

REFERENCES

BERLEW, D. W., & HALL, T. E. The socialization of managers. The effects of expectations on performance. *Administrative Science Quarterly*, 1966, **11**, 207–223.

COHEN, J. Multiple regression as a general data-analytic system. *Psychological Bulletin*, 1968, **70**, 426–443.

FARRIS, G. F. Organizational factors and individual performance: A longitudinal study. *Journal of Applied Psychology*, 1969, **53**, 87–92.

HEMPHILL, J. K. (Ed.) *The engineering study*. Princeton, N.J.: Educational Testing Service, 1963.

KAUFMAN, H. G. Relations of ability and interest to currency of professional knowledge among engineers. *Journal of Applied Psychology*, 1972, **56**, 495–499.

KAUFMAN, H. G. A convergent and discriminant validity analysis of utilization measures among engineers. *Studies in Personnel Psychology*, 1973, **5**, 75–80.

KAUFMAN, H. G. Obsolescence and professional career development. New York: American Management Association, 1974, in press.

LEBOLD, W. K., PERRUCCI, R., & HOWLAND, W. The engineer in industry and government. *Journal of Engineering Education*, 1966, **56**, 237–259.

RONEY, J. G. *Report of first conference on occupational obsolescence* (Dallas, Texas, June 21–22, 1966). Menlo Park, Calif.: Stanford Research Institute, 1966.

(Received March 12, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 380-382

SOME MODERATOR INFLUENCES ON RELATIONSHIPS BETWEEN CONSIDERATION, INITIATING STRUCTURE, AND ORGANIZATIONAL CRITERIA[1]

IRWIN J. BADIN[2]

New York University

Group size, tenure, position power, and task structure were studied as potential moderators of the relationships between Consideration (C) and Initiating Structure (IS) and employee satisfaction and group effectiveness. The subjects were 489 male employees of a large manufacturing company. C was positively related to satisfaction across all moderator conditions and to group effectiveness in high-position power and high-task structure conditions. IS was negatively related to group effectiveness for small groups, seasoned employees, low-position power, and high-task structure. There was no significant relationship between IS and either criterion for large groups, new employees, high-position power, and low-task structure. The consistency of the C results as compared to the more situational nature of the IS result, replicated other findings.

The identification and description of the Consideration (C) and the Initiating Structure (IS) dimensions of leader behavior (Fleishman, 1953) has served as an important framework for much of the contemporary research into leadership effectiveness. However, several authors have pointed out the need for better definitions of the conditions that may moderate the relationships between these leader behaviors and various measures of organizational effectiveness (Badin, 1971; Fleishman, 1973; Fleishman, Harris, & Burtt, 1955; Korman, 1966; Vroom, 1973).

The aim of the present research was to examine the extent to which the relationships between C and IS and measures of employee satisfaction and group effectiveness were moderated by certain situational characteristics. Four potential moderators, commonly hypothesized sources of influence in leadership theory, were chosen; work group size, subordinates' tenure, supervisors' position power, and extent of task structure.

Work Group Size

Generally, it has been found that employees tend to accept leader-centered direction in large work groups (Hemphill, 1950; Vroom & Mann, 1960). As for C, a guiding rationale throughout the study was that people-oriented behavior should be related to effectiveness and satisfaction regardless of variations within any of the moderators. Accordingly, the following hypotheses were formulated: (*a*) The correlations between IS and the criteria are positive within large groups and negative within small groups. (*b*) The correlations between C and the criteria are positive within large and small groups.

Tenure

Bass (1960) points out that structuring and organizing leader behaviors help the new employee understand his environment. A continuing emphasis on structuring the role of a seasoned employee, however, leads to his dissatisfaction. Accordingly, the following hypotheses were formulated: (*a*) The correlations between IS and the criteria are positive for new employees and negative for seasoned employees. (*b*) The correlations between C and the criteria are positive for new and seasoned employees.

Position Power

Fiedler (1967) has stated that it is more "favorable" to the leader if he has high-position power and that subordinates are more likely to accept task-oriented behavior from a manager who was perceived to have such high-position power than low-position power. Accordingly, the following hypotheses were formulated: (*a*) The correlations between IS and the criteria are positive for employees who perceive that their managers have high-position power and is negative for those who perceive that their managers have low-position power. (*b*) The correlations between

---

[1] This study is based on a doctoral dissertation submitted to the Department of Psychology, New York University. The author gratefully acknowledges the helpful suggestions of Abraham K. Korman and Jeffery H. Greenhaus. The author is also indebted to John R. Hinrichs of International Business Machines for his cooperation and assistance.

[2] Requests for reprints should be sent to the author who is now at Montclair State College, Department of Psychology, Upper Montclair, New Jersey 07043.

TABLE 1

CORRELATIONS BETWEEN LEADER BEHAVIOR AND EFFECTIVENESS ACROSS DIFFERENT CONDITIONS OF SITUATIONAL VARIABLES (ns IN PARENTHESES)

Moderator	Correlation			
	IS–sat	Group $\bar{X}$ IS–group effectiveness	C–sat	Group $\bar{X}$ C–group effectiveness
Group size				
Large ($n \geq 12$)	−.04 (254)	−.02 (19)	.33** (254)	.34 (19)
z_{diff}	1.85	2.00*	1.05	.23
Small ($n \leq 11$)	.13* (227)	−.60** (23)	.22** (228)	.27 (23)
Tenure				
New (≤6 months)	.00 (185)	−.19 (21)	.33** (185)	.31 (21)
z_{diff}	.74	.80	1.16	.00
Seasoned (>6 months)	.07 (206)	−.43* (21)	.23** (297)	.31 (21)
Position power				
High	.12 (201)	−.15 (18)	.19** (201)	.52* (18)
z_{diff}	1.69	1.10	.39	1.33
Low	−.05 (197)	−.51* (17)	.23** (197)	.08 (17)
Task structure				
High	−.03 (148)	−.56* (17)	.22** (148)	.55* (17)
z_{diff}	.43	2.17*	.26	.99
Low	.02 (153)	.20 (16)	.25** (153)	.23 (16)

Note. IS = initiating structure; Sat = individual satisfaction, C = consideration.
* $p < .05$.
** $p < .01$.

C and the criteria are positive regardless of the leader's perceived position power.

Task Structure

Fiedler (1967) has posited that it is more favorable to the leader if there is a high degree of task structure. An important question is whether high-task structure can lead to effectiveness regardless of the leader's behavioral orientation. It seems likely that in conditions in which the task is inherently structured, any further attempts on the manager's part to initiate structure would be seen as superfluous and would interfere with employee effectiveness. For tasks which are perceived to be ambiguous, structuring behavior may be seen as more beneficial. Accordingly, the following hypotheses were formulated: (*a*) The correlations between IS and the criteria are negative for employees who perceive their task as being high in structure and positive for those who perceive it as being low in structure. (*b*) The correlations between C and the criteria are positive regardless of the variation in perceived task structure.

METHOD

Subjects

Subjects were 489 male employees of a large manufacturing company, who performed their duties in one of 42 work groups. All subjects were in some way directly involved in the marketing efforts of the company.

Research Instruments

Leader behavior variables. C and IS scores were obtained from items that appear in the Supervisory Behavior Questionnaire (Fleishman, 1957). Although this questionnaire has 48 items, only the 24 items with the highest and purest loading patterns were chosen for the present research (12 items for each dimension).

Moderator variables. Group size was obtained from company records. Tenure was obtained from a question in the survey booklet. The extent to which the manager was perceived to have a position of power was based on the mean of five items which were completed by each subordinate in his group. The content of the items were obtained from a "Position Power Questionnaire" devised by Fiedler (1967). Task structure has been defined by four factors (Shaw, 1967): decision verifiability, goal clarity, goal-path multiplicity, and solution specificity. The task structure score was based on the mean of four items, one for each factor.

Criterion variables. Individual satisfaction was measured by a 6-item scale which had been used in several previous job satisfaction surveys at the company. The group's effectiveness was measured by a 12-item rating scale which was completed by the branch office manager to whom the supervisor of each work group reported.

Questionnaires were filled in anonymously, coded according to supervisor, and mailed to an office at New York University.

RESULTS AND DISCUSSION

Before undertaking the hypothesized moderator analyses, correlations were computed for the different leader behaviors over all conditions. This analysis indicated positive relationships for the C dimension for both the individual satisfaction ($r = .26$, $p < .01$, $n = 484$) and group effectiveness ($r = .35$, $p < .05$, $n = 42$) criteria. On the other hand, IS showed no significant relationship to individual satisfaction ($r = .04$, $n = 484$) and a negative relationship to group effectiveness ($r = -.36$, $p < .05$, $n = 42$). Hence, some relationships did exist on the general level, relationships which it was hoped the moderator analyses would clarify further.

Table 1 presents the data for these predicted effects. (A median split on the moderator was used in all cases except tenure which was dichotomized on an a priori basis). As predicted, IS was negatively related to group effectiveness for small groups, seasoned employees, low-position-power managers, and high-task-structure conditions. These relationships were all significant. However, contrary to the hypotheses, no significant positive relationships were found. Group size and task structure moderated the relationship between IS and group effectiveness. No significant relationships were found between IS and individual satisfaction in any of the conditions except small group size ($r = .13$, $p < .05$).

All of the correlations between C and individual satisfaction were positive and significant at the .01 level. For ratings of group effectiveness, all correlations were positive, although significant in only two cases (high-position power, high-task structure).

The positive relationships between C and the criteria over all conditions remained quite stable throughout all levels of the moderator variables. Although the overall negative correlation between IS and group effectiveness was found in certain levels of the moderators (those predicted to contain significant negative relationships), in others (those predicted to contain significant positive relationships), no significant relationships existed. The finding that the C results were more consistent across situations than were the IS results was consonant with the prevailing trend of data reported in the literature (Fleishman, 1973). The findings of this type of research should help specify the relative importance of specific moderators as well as those instances in which situational variation may not influence the relationship between leader behavior and effectiveness. This need for such specification has been stressed by Fleishman (1973) and Korman (1973).

REFERENCES

BADIN, I. J. Some moderator influences on the relationship between consideration, initiating structure and organizational criteria. Unpublished doctoral dissertation, New York University, [illegible].

BASS, B. M. *Leadership, psychology, and organization behavior.* New York: Harper, 1960.

FIEDLER, F. E. *A theory of leadership effectiveness.* New York: McGraw-Hill, 1967.

FLEISHMAN, E. A. The description of supervisory behavior. *Journal of Applied Psychology,* 1953, 37, 1–6.

FLEISHMAN, E. A leader behavior description for industry. In R. M. Stogdill & A. E. Coons (Eds.), *Leader behavior: Its description and measurement.* Columbus: Ohio State University, Bureau of Business Research, 1957, (Research Monograph No. 88) 103–119.

FLEISHMAN, E. A. Twenty years of consideration and structure. In E. A. Fleishman & J. G. Hunt (Eds.), *Current developments in the study of leadership.* Carbondale: Southern Illinois University Press, 1973.

FLEISHMAN, E. A., HARRIS, E. F., & BURTT, H. E. *Leadership and supervision in industry.* Columbus: Ohio State University, Bureau of Educational Research, 1955.

HEMPHILL, J. K. Relations between the size of the group and behavior of the superior leaders. *Journal of Social Psychology,* 1950, 32, 11–22.

KORMAN, A. K. "Consideration," "initiating structure" and organizational criteria—A review. *Personnel Psychology,* 1966, 19, 349–361.

KORMAN, A. K. On the development of contingency theories of leadership: Some methodological considerations and a possible alternative. *Journal of Applied Psychology,* 1973, 58, 384–387.

SHAW, M. E. Scaling group tasks: A method for dimensional analysis. In F. E. Fiedler, *A theory of leadership effectiveness.* New York: McGraw-Hill, 1967.

VROOM, V. *Work and motivation.* New York: Wiley, 1964.

VROOM, V., & MANN, F. C. Leader authoritarianism and employee attitudes. *Personnel Psychology,* 1960, 13, 125–140.

(Received May 1, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 383–385

A NOTE ON THE PREDICTIVE VALIDITY OF INITIATING STRUCTURE AND CONSIDERATION IN ROTC TRAINING

M. M. PETTY[1]

Manpower and Industrial Relations Institute
University of Alabama

NORMAN M. PRYOR[2]

University of Tennessee

Sixty-eight advanced Reserve Officer Training Corps (ROTC) students participated in a Leaderless Group Discussion (LGD), filled out the Leadership Opinion Questionnaire (LOQ), and were described by peers with Form 12 of the Leadership Behavior Description Questionnaire (LBDQ). Four to five months later, measures of their performance during summer field training with the U.S. Army were obtained and correlated with their scores on Initiating Structure (IS) and Consideration (C) as measured by the LGD, LOQ, and LBDQ. LGD-IS and LGD-C were significantly correlated with several of the performance measures. LGD-IS was the best predictor ($r = .45$) of the Leadership Potential Index (LPI), the Army's overall rating, while LBDQ-C was significantly correlated with the Field Problem Test.

Much of the evidence on the validity of the two well-known dimensions of leadership behavior, Initiating Structure (IS) and Consideration (C), is concurrent validity. Korman (1966) and Fleishman (1972, 1973) have pointed out the need for more studies of the predictive validity of measures of these dimensions.

These leadership dimensions have been measured by several instruments, most notably various forms of the Leadership Behavior Description Questionnaire (LBDQ), by means of which subordinates typically describe their supervisor (Stogdill & Coons, 1957; Fleishman, 1953a, 1972); and the self-report Leadership Opinion Questionnaire (LOQ), (Fleishman, 1953b, 1969). The Leaderless Group Discussion (LGD) is another procedure that has been used to measure IS and C (Bass, 1954). In the case of the LBDQ and LOQ the leadership dimensions and scoring keys were identified by psychometric techniques such as item analysis and factor analysis. The LGD measures are inferred from an overall single behavior rating provided by observers of the group members activity. Earlier, Bass had reported eight predictive validities of .33–.65, with a median of .46 for the LGD-IS score.

The purpose of the present study is to examine the predictive validities of IS and C as measured by the LOQ, LBDQ, and LGD, for a sample of ROTC students, using several criteria of leader performance.

METHOD

Subjects

The subjects were 68 ROTC students at the University of Tennessee, beginning their third year of training in military science.

Procedure

Subjects were administered the LOQ in class. During the same class period each subject participated in a four-man LGD from which measures of IS and C (Bass, 1954, p. 468) were obtained. Each subject was rated by two senior ROTC students who served as laboratory assistants to the instructor. Seven weeks later each subject was asked to describe three of his peers[3] with the 100 item, Form 12 of the LBDQ. The IS and C scales each contained 10 items. Four to five months later the subjects attended six weeks of summer field training at a U.S. Army fort. During the course of their training the subjects were evaluated on the following criteria: (*a*) three graphic ratings by the training officer including an assessment of the subject's leadership abilities in four, one-day leadership roles, personal characteristics of the subject (e.g., "knows how to handle men"), and the overall willingness of the training officer to allow the subject to represent him in any situation; (*b*) peers' rating (PEER) of their willingness to follow the subject into combat; (*c*) Field Problem Test (FIELD)—an average behavioral

---

[1] Requests for reprints should be sent to M. M. Petty, P. O. Box 2737, University of Alabama, Alabama 35486.

[2] Now at Ford Motor Company, Dearborn, Michigan.

[3] The ratings were more than peer ratings, since all students had to assume "leadership positions" (platoon leader, squad leader, etc.) during the course of their training. Thus at various times, the raters were subordinates of their "peers."

TABLE 1

INTERCORRELATIONS BASED ON 68 FIELD TRAINING PARTICIPANTS

Measure	IS-LOQ	C-LOQ	IS-LGD	C-LGD	IS-LBDQ	C-LBDQ	PEER[a]	FIELD[b]	LPI[c]
LOQ-IS									
LOQ-C	−.17								
LGD-IS	.09	.07							
LGD-C	−.02	.02	.73**						
LBDQ-IS	−.02	.04	.27*	.22					
LBDQ-C	−.05	.08	.12	.04	.36**				
PEER	.24	.11	.42**	.20	.05	.12			
FIELD	−.14	.18	.10	−.03	.10	.31**	.38**		
LPI	.09	.10	.45**	.28*	.08	.22	.77**	[illegible]	

Note. LOQ = Leadership Opinion Questionnaire; IS = Initiating Structure; LGD = Leaderless Group Discussion; C = Consideration; LBDQ = Leadership Behavior Description Questionnaire; PEER = peers' rating; FIELD = Field Problem Test; LPI = Leadership Potential Index.

[a] Peers' rating.

[b] Field Problem Test.

[c] Leadership Potential Index.

* $p < .05$.

** $p < .01$.

rating by a noncommissioned officer of the subject's leadership ability in three or four simulated military missions; and (*d*) cadet Leader Potential Index (LPI)—a weighted composite of all[4] summer field training criteria.

RESULTS

The intercorrelations between IS and C measures and the performance measures may be found in Table 1. The graphic ratings provided by the training officer have been omitted from this table because of their high intercorrelations (.83–.96) and their high correlations with the LPI (.88–.95). The peers' rating and the Field Problem Test are included in Table 1 even though they are contained in the LPI, since some degree of independence is demonstrated and they are conceptually separable.

An examination of Table 1 indicates that LGD-IS appears to be the best predictor of the peers' rating and the LPI, while LBDQ-C was the only significant predictor of the Field Problem Test. LGD-C is also significantly correlated with the LPI, but its addition to LGD-IS in a multiple regression equation does not add any incremental predictive validity for the LPI. The LOQ failed to predict any of the performance criteria.

The low heteroinstrument–homeotrait correlations appear to indicate that the three instruments were measuring different phenomena (i.e., self, peer, and observer perceptions of leadership status). The low-homeoinstrument-heterotrait correlations for the LOQ and LBDQ are consistent with the previous findings reported by Fleishman (1969, 1972). The high correlation between LGD-IS and LGD-C is similar to the results of Bass (1954) who believes that the LGD is more a measure of IS than C.

DISCUSSION

The results of the present study indicated that LGD-IS was the best predictor of the most important leadership measure, the LPI. This latter measure is important because it is used to make some personnel decisions that will affect the subject's future with the Army (e.g., branch assignments). One possible explanation for the superior predictive power of LGD-IS was that the conditions were very similar under which both the LGD measures and many of the summer camp measures were obtained. That is, the subject's behavior was observed and evaluated immediately by independent observers. The superior predictive power of the LGD in this study appears to lend support for the continued use of the LGD in asssessment centers.

The generality of results from this study is limited. Some predictive validity of IS and C is demonstrated, but this depends on the criteria and the instruments used.

REFERENCES

BASS, B. M. The leaderless group discussion. *Psychological Bulletin,* 1954, **51**, 465–492.

---

[4] There were four other measures of summer field training performance in addition to the three cited in text. The additional criteria included: rifle marksmanship, physical fitness, map and compass reading ability, and a comprehensive test covering the cognitive aspects of summer field training.

FLEISHMAN, E. A. The description of supervisory behavior. *Journal of Applied Psychology*, 1953, **37**, 1-6. [illegible]

FLEISHMAN, E. A. The measurement of leadership attitudes in industry. *Journal of Applied Psychology*, [illegible] 37, 153-158. (b)

FLEISHMAN, E. A. *Manual for the Leadership Opinion Questionnaire*. Chicago: Science Research Associates, 1969.

FLEISHMAN, E. A. *Manual for the Supervisory Behavior Description Questionnaire*. Washington: American Institutes for Research, 1972.

FLEISHMAN, E. A. Twenty years of consideration and structure. In E. A. Fleishman and J. G. Hunt (Eds.), *Current problems in the study of leadership*. Carbondale: Southern Illinois University Press, 1973.

KORMAN, A. K. "Consideration," "initiating structure," and organizational criteria—A review. *Personnel Psychology*, 1966, **19**, 346-361.

STOGDILL, R. M., & COONS, A. E. (Eds.) *Leader behavior: Its description and measurement.* Columbus: Ohio State University, Bureau of Business Research, 1957.

(Received April 24, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 386–388

RELATIONSHIP BETWEEN SUPERVISORY COGNITIVE STYLE AND SOCIAL ORIENTATION

L. W. GRUENFELD[1]
Cornell University

P. WEISSENBERG
State University of New York at Binghamton

Supervisors from three different organizations completed a version of the Embedded Figures Test and ranked 10 inner/other-directed adjectives according to their importance for success on their jobs. The hypothesis that field independent supervisors rank inner-directed attributes as most important for success was clearly supported for the attribute of "independence" in all three organizations. For the remaining attributes differential effects depending on the type of organization were observed, but the direction of the observed relationships was consistent with the hypothesis.

According to Witkin, Dyk, Faterson, Goodenough, and Karp (1962) when a person interacts with his environment, his responses vary along a continuum from articulated (field independent) to global (field dependent). At the field independent end of the continuum, the characteristic social behavior is independent and self-reliant. Linton (1955) has found empirical support for this relationship between field independence and independent social behavior. Using Asch's classical conformity studies, Linton observed that field independent individuals were much more able to resist social pressure for conformity. Cattell (1971) has added further evidence supporting this relationship. He noted that measures of field independence are a manifestation of the source trait "temperamental independence" in the perceptual performance realm. Temperamental independents or field independents showed self-assured and independent behaviors.

The descriptions of the behavioral referents of field independence given by Witkin and Cattell overlap considerably with Riesman's (1950) description of inner-directed types. Riesman indicated that individuals who were inner-directed were secure, relatively self-assured, and socially independent. Taken together, these findings suggest that there is a direct relationship between field independence as measured in the perceptual realm and inner-directedness as measured in terms of preferred social orientation.

At the other end of the continuum of responses to the environment, the field dependent end of the continuum, the characteristic social behavior is dependent and other-directed. This finding by Witkin et al. (1962) suggests that there is a direct relationship between field dependence as measured in the perceptual realm of preferred social orientation. Riesman (1950) indicated that individuals who were other-directed emphasized the need for security, approval, and social dependency.

Based on the presumption that individuals perform their roles in a manner congruent with the behavioral pattern which is characteristic of their cognitive style, the following specific hypothesis was formulated to guide this research: When asked to rank traits which they consider important for success on their present jobs, field independent supervisors consistently rank inner-directed traits as more important, whereas field dependent supervisors consistently rank other-directed traits as more important.

METHOD

Sample

The subjects came from three different organizations. The first group, a production group consisted of 73 male lower- and middle-level managers in a food processing plant. The second group included 51 male managerial employees of a state department of taxation and finance, and the third group consisted of 35 male and female managerial and professional employees of a county government unit. The county government group was relatively field dependent; the state civil service and production groups were more representative although the latter was relatively more field independent than the other two groups.

Research Instruments

Three different but related instruments were used to measure field independence. The individual Embedded Figures Test, EFT (Jackson, 1956), was administered to the state civil service sample. The Hidden Figures Test, HFT-V (Jackson, Messick, & Myers, 1964), a group form, was administered to the production sample, and the Hidden Figures Test, Cf-1 (French, Ekstrom, & Price, 1963), was administered in group meetings to the county government group. All three of these instruments are essentially similar in that they require the subject to find a simple figure in an embedding background. Jackson et al. (1964) and Weissenberg (1973) have indicated that all three forms are valid measures of psychological differentiation.

---

[1] Requests for reprints should be sent to L. W. Gruenfeld, New York State School of Industrial and Labor Relations, Cornell University, Ithaca, New York 14850.

The instrument used to measure the importance of inner/other-directed traits for success on the present job was developed by Porter and Lawler (1968). A list of 10 adjectives is presented to the subject and he is instructed to rank them in the "order of their importance for success in [his] present management position." The most important trait is ranked as "1," while the least important is ranked as "10." The inner-directed traits listed are forceful, imaginative, independent, self-confident, and decisive. The other-directed traits are cooperative, adaptable, cautious, agreeable, and tactful.

RESULTS AND DISCUSSION

Table 1 shows the significant Pearson product-moment correlations between rankings of certain of the traits and scores on a measure of field independence.

In the case of one adjective, independent, there was clear support for the hypothesis in all three groups. The field independent individual ranks this inner-directed trait as being more important for success than does the field dependent. For the imaginative trait only the production group results supported the hypothesis. No significant correlations were observed between rankings of the other three inner-directed adjectives: forceful, self-confident, and decisive, and scores on the various measures of field independence/field dependence.

For the other-directed adjectives, field dependent subjects in the production and county government groups ranked tactful as being more important for success. Positive correlations between rankings of the two other-directed traits, cautious and agreeable, and scores on the Hidden Figures Test in the county government group further indicated that field dependent individuals ranked these two traits as more important for success. In no group were there significant correlations between rankings for the remaining other-directed adjectives, adaptable and cooperative, and scores on the measure of field independence/field dependence.

TABLE 1

CORRELATIONS BETWEEN TRAIT RANKINGS AND SCORES ON A MEASURE OF FIELD INDEPENDENCE

Adjective	Production group ($n = 73$)	State civil service group ($n = 50$)	County Government group ($n = 35$)
Imaginative	−.25*	—	—
Independent	−.25*	−.34*	−.38*
Tactful	.31*	—	.46**
Cautious	—	—	.37**
Agreeable	—	—	.50**

* $p < .05$.
** $p < .01$.

Table 1 reveals (*a*) a strong predominance of correlations between rankings of inner-directed traits and scores on the measure of field independence/field dependence in the production group and (*b*) a strong predominance of correlations between ranks of the other-directed traits and the measure of field independence/field dependence in the county government group. In the state civil service group there is only one correlation: that between the ranking of independence and the score on the individual EFT. All correlations observed are in the expected direction, but further research needs to examine the fact that there were significant correlations for only some of the characteristics and that these do not occur in all three groups.

Nevertheless, as Table 1 shows, there is clear evidence that field independent managers, regardless of the type of organization, rank independence higher than do field dependent supervisors as an attribute required for managerial success. These results provide evidence for the validity of the construct field independence as measured by various forms of the Embedded Figures Test. They further support the proposition that Riesman's inner-directedness and Witkin's field independence are related. Furthermore, the field independent/field dependent cognitive style is related to characterictics believed to be associated with managerial success.

There are several additional implications for further research suggested by these findings. Porter and Lawler (1968) related differences in the perception of traits required for success to the effectiveness of supervisory performance.

Inner-directed supervisors were ranked higher on job effectiveness than were other-directed supervisors. In addition, in an unpublished study the authors found that in a manufacturing organization first- and second-line supervisors who were field independent were ranked by their supervisors as more successful than field dependent supervisors ($r = .48$, $p < .01$). When combined with the results of the present study, these findings suggest that measures of field independence may be fruitfully used in subsequent studies of leadership effectiveness.

REFERENCE

CATTELL, R. B. *Abilities: Their structure, growth, and action.* Boston: Houghton Mifflin, 1971.

FRENCH, J. W., EKSTROM, R. B., & PRICE, L. A. *Manual for kit of reference tests for cognitive function.* (Rev. ed.) Princeton, N.J.: Educational Testing Service, 1963.

JACKSON, D. N. A short form of Witkin's Embedded-Figures Test. *Journal of Abnormal and Social Psychology,* 1956, **53**, 254–255.

JACKSON, D. N., MESSICK, S., & MYERS, C. Evaluation of group and individual forms of embedded-figures measurements of field-independence. *Educational & Psychological Measurement,* 1964, **24**, 177–192.

LINTON, H. B. Dependence on external influence: Correlates in perception, attitudes and judgment. *Journal of Abnormal and Social Psychology,* 1955, **51**, 502–507.

PORTER, L. W., & LAWLER, E. E. *Managerial attitudes and performance.* Homewood, Illinois: Irwin, 1968.

RIESMAN, D. *The lonely crowd.* Garden City, N.Y.: Doubleday, 1950 (unabridged).

WEISSENBERG, P. Concurrent validity of the Hidden Figures Test (Cf–1). *Perceptual and Motor Skills,* 1973, **36**, 460–462.

WITKIN, H. A., DYK, R., FATERSON, H., GOODENOUGH, D., & KARP, S. *Psychological differentiation.* New York: Wiley, 1962.

(Received April 4, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 389–391

DOGMATISM AND THE CONGRUENCE BETWEEN SELF-REPORTED JOB PREFERENCE AND PERFORMANCE AMONG SCHOOL SUPERVISORS

JAMES P. ESPOSITO[1] AND HERBERT C. RICHARDS

University of Virginia

One hundred and seventy-four instructional supervisors were administered the Rokeach Dogmatism Scale and an instrument requiring the ranking of 10 supervision tasks according to ideal and actual priorities. Respondents were grouped into three categories according to level of measured dogmatism. Findings indicated that closed-minded individuals manifest significantly lower correspondence of ideal and actual task rankings than did their medium- or low-dogmatic counterparts.

One of the most persistent findings among studies of supervisory personnel and their multifaceted role in public education is the existence of major discrepancies between what supervisors actually do during the work day and tasks they believe most important or desirable to perform (e.g., Carman, 1970). Although a variety of explanations have been suggested for the surprisingly poor match of intention and performance, findings of the Regional Curriculum Project (1968) indicate that supervisors perceive a multitude of administrative or political hindrances (real or imaginary) to job performance. Undoubtedly there are some functional hindrances to performing a supervisory job such as filling out routine forms or the preparation of bureaucratic reports, but such hindrances plague most jobs performed in an organizational setting. It is equally apparent that some individuals, despite the limitations imposed by the "system," adequately perform those tasks they rate as most interesting and important. The presence of individual variability implies that something in the personal-social makeup of the individual supervisor lies at the core of his inability to match intention with performance rather than in the external matrix of bureaucratic hindrances. Thus, both personal competence and individual temperament are implicated as key antecedent variables.

One of the most salient aspects of an individual's temperament is the extent to which his core beliefs about the world are flexible, open to modification, and subject to change based on incoming information. In fact, since the concept of belief system rigidity was first subjected to quantitative measurement under the rubric of authoritarianism (Adorno, Frenkel-Brunswick, Levinson, & Sanford, 1950), psychologists have identified its manifestations through a multitude of criterion variables. However, the dependence of closed-minded individuals on authority figures (e.g., Feather, 1967) provides the basis for the present study which involves supervisory personnel. More specifically, the purpose of this study is to investigate the prediction that out of deference to authority figures closed-minded supervisors will perform those tasks prescribed for them by their superiors regardless of personal priorities; and furthermore, that closed-minded individuals will manifest a lower correspondence between ideal and actual task self-rankings than their open-minded colleagues.

METHOD

Subjects

The invited sample (those initially contacted as potential subjects) consisted of 274 individuals from 67 Florida counties who were identified as general supervisors, elementary and secondary consultants and specialists, or area instructional coordinators. Directors of instruction and assistant directors of instruction were included for those counties where no other general supervisors were listed in the Florida Educational Directory. Although data were obtained for 203 subjects (approximately 74% of the invited sample), 29 questionnaires were discarded because of incompleteness or because the respondent chose not to participate. Thus, the effective sample contained 174 subjects, 64% of those initially identified for sampling.

Measures

The Rokeach (1960) Dogmatism Scale (Form E) was adopted as the independent measure for the present investigation. This instrument yields a three-digit score intended to indicate the degree to which the belief system of an individual is open to change on the basis of new experience. The Rokeach Dogmatism Scale has an impressive validity network, and readers are referred to Rokeach (1960) for an extended (although somewhat dated) treatment of this topic.

---

[1] Requests for reprints should be sent to James P. Esposito, Curry Memorial School of Education, University of Virginia, Charlottesville, Virginia 22903.

The criterion measures for the present study consist of forced choice rankings of the 10 major tasks of supervision identified by Harris (1963): developing curriculum; organizing for instruction; staffing; providing facilities; providing materials; arranging for in-service education; orienting new staff members; relating special services; developing public relations; evaluating. Each subject was presented with a listing of these tasks and a description of each. The ranking procedure was explained through the following written instructions:

> Please rank order from 1 through 10, the 10 tasks in terms of the frequency in which you (1) actually perform that task (actual), and (2) would prefer to perform that task if given the opportunity (ideal).
>
> Indicate the actual rank in the space provided at the left of each task and the preferred or ideal rank at the right of each task. The ideal rank should be determined by *your preference* as to the frequency in which you would prefer to perform each task.
>
> Place a *1* by the task with the highest frequency of performance, *2* by the task with the second highest frequency, *3* by the task with the third highest frequency . . . *9* by the task with the ninth highest frequency and *10* by the task with the lowest frequency.

Thus, in addition to dogmatism, 20 scores vere obtained from each participant, 10 correponding to actual task rankings and 10 others orresponding to ideal task rankings.

Analysis

In order to classify subjects according to level of measured dogmatism, all respondents who cored in the upper third of the sample of the Rokeach scale were designated as high dogmatics ($n = 57$), while those who scored in the lower hird were designated as low dogmatics ($n = 56$). Subgroup sample sizes were not perfectly equal due to ties on the dogmatism scale). Thus, in ddition to having a three-digit dogmatism score, each subject was classified as high, medium, or low on dogmatism.

To facilitate the interpretation and reporting of the data, all task ranks were inverted directionally so that a high score indicated high preference for (or performing of) the task. This was done by subtracting each ranking from a constant (11). The resulting scale ranged from 1 to 10; a 1 indicated low preference or performance.

The correspondence of actual and ideal task rankings was estimated for each subject by computing a rank-order correlation (i.e., Spearman's Rho) between the two sets of scores. Differences between real and ideal rankings for each task separately were also obtained by subtracting ideal from actual ranks.

Mean Spearman rhos for high-, medium-, and low-dogmatic groups were computed and differences among these means across the three groups were tested for significance through analysis of variance (independent group design). Differences between individual group means (pair-wise comparisons) were tested for significance by the Newman-Keuls method (see Winer, 1962). To identify more specifically the task rankings that are differentially associated with dogmatism, actual rankings, ideal rankings, and differences between actual and ideal rankings were each correlated with dogmatism.

RESULTS AND DISCUSSION

The results of the overall analysis of variance were significant ($F = 12.2$, $p < .001$). Thus, level of dogmatism is related to the correspondence of actual and ideal task rankings. However, inspection of the individual group means (.34, .63, and .59 for high, medium, and low groups, respectively) reveals the effect to be a nonlinear one. Whereas high dogmatics manifested significantly lower Spearman rhos than either of their medium ($q_3 = 6.5$, $p < .01$) or low ($q_2 = 5.6$, $p < .01$) dogmatic counterparts, the difference between medium and low dogmatics was not significant ($q_2 = .91$, ns).

Significant product-moment correlations of dogmatism with differences in actual-ideal rankings were found for Task 1 (developing curriculum; $r = -.52$) and Task 2 (organizing for instruction; $r = -.35$). Other significant correlates of dogmatism include actual rankings of Task 1 ($r = -.61$) and ideal rankings of Task *2* ($r = .45$). It is apparent from these results that high dogmatism is associated with low-actual rankings of curriculum development and high-ideal rankings of organizing for instruction.

These data clearly support the hypothesis of the present investigation and help augment the construct validity of the Rokeach Dogmatism Scale. Although limited by the self-report nature of the data, the finding that closed-minded supervisors manifest significantly lower ideal-actual correspondences of task priorities than do more open-minded colleagues is consistent with the view that closed-minded individuals are more subject to the external control of authority figures. Furthermore, the correlations of dogmatism with task rankings support the notion that highly dogmatic individuals see themselves as performing innovative tasks (e.g., develop curricula) less, and they idealize task organization (e.g., organize for instruction) more than do their low-dogmatic counterparts. Taken together, these results lend credence to the theoretical formulations of Rokeach (1960) and help account for the puzzling variability among supervisors in the degree to which they match priorities and perceived performance. If the correspondence of ideal and actual task rankings is viewed as a sensible index of job satisfaction, it also follows that dogmatic supervisors are generally less content with their work. However, it must be emphasized that the present results are exploratory and additional research is needed to validate the link between indices of job satisfaction and the correspondence of task rankings. Hopefully, such studies provide additional insights into how personality influences job satisfaction and performance.

REFERENCES

ADORNO, T. W., FRENKEL-BRUNSWICK, E., LEVINSON, D. J., & SANFORD, R. W. *The authoritarian personality.* New York: Harper, 1950.

CARMAN, B. D. Roles and responsibilities of general supervision of instruction: a synthesis of research findings. Unpublished doctoral dissertation, Florida State University, 1970.

FEATHER, N. T. Evaluation of religious and neutral arguments in religious and atheist student groups. *Australian Journal of Psychology,* 1967, **19,** 3–12.

HARRIS, B. M. *Supervisory behavior in education.* Englewood Cliffs, N.J.: Prentice-Hall, 1963.

REGIONAL CURRICULUM PROJECT (RCP). A role study: The county level supervisor in Florida. Atlanta, Ga.: RCP, Trinity Avenue, S.W., 1968.

ROKEACH, M. E. *The open and closed mind.* New York: Basic Books, 1960.

WINER, B. J. *Statistical principles in experimental design.* New York: McGraw-Hill, 1962.

(Received January 31, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 392-394

UNITY OF COMMAND AND JOB ATTITUDES OF MANAGERS IN A BUREAUCRATIC ORGANIZATION[1]

MARTIN J. GANNON[2] AND FRANK T. PAINE[3]

Department of Business Administration, University of Maryland

Several principles, commonly termed the classical principles of management, although enunciated in order to help managers increase organizational efficiency and effectiveness, have been tested only on a minimal basis. The present study concerns the classical principle of unity of command, that is, a subordinate should report only to one superior. The subjects were 304 government managers working in a bureaucratic agency, 181 of whom reported only to one superior while 123 reported to two or more superiors. Results indicated that, at least within a bureaucratic milieu, the violation of this principle was related to the predicted negative attitudinal orientation of managers.

An important development in the field of organizational psychology within recent years has been the empirical testing of propositions related to the classical school of management thought (see Filley & House, 1969). Such testing is critical if the classical principles of management thought are to be applied in a precise and beneficial manner.

One of the prominent classical principles is unity of command. This principle postulates that a subordinate should receive orders only from one superior. Only a few research studies of this principle have been completed. A recent review of the research has suggested that the violation of the principle does, in fact, result in dysfunctional consequences (Rizzo, House, & Lirtzman, 1970; see also Kahn, Wolfe, Quinn, & Snoek, 1964). These consequences include an increase in role conflict, employee dissatisfaction, and ineffective performance.

However, some theorists have argued that the strict use of such classical principles appears to maximize efficiency and reduce role conflict only when the technology demands a bureaucratic organizational structure (e.g., see Lawrence & Lorsch, 1967). Such a situation would be approximated by a technology of medium complexity, for example, mass production assembly. As the technology becomes more complex, it appears that individuals are more satisfied if more flexibility is provided through the violation of the principle of unity of command, since the requirements of the work necessitate such a system. This particular situation is best approximated in technologically advanced process industries such as chemicals.

Some descriptive research appears to substantiate this point of view. For instance, Woodward (1965) found that supervisors and managers in process industries generally enjoyed their work when the principle of unity of command was openly violated, since they needed flexibility in order to carry out their assignments. Conversely, individuals responded favorably to the strict use of the principle of unity of command when they worked in an organization that was structured in terms of medium technological complexity, for example, the automotive assembly line.

The present study was conducted in a large government agency, the General Services Administration, which is a large nationwide organization engaged in supply, maintenance, and protective activities. In terms of Thompson's (1967) well-known typology, this agency can be characterized as one with a mediating technology, since its basic purpose is to link clients or customers who are interdependent. Under such conditions the imposition of bureaucratic rules of behavior appears to be most advantageous. Consequently, we expect the strict use of the principle of unity of command to be associated with less role conflict and increased employee effectiveness within this bureaucratic setting. In effect, these are the major hypotheses under investigation in the present study.

METHOD

Owing to the complexity and geographical dispersion of operations in General Services Administration, the researchers decided to use a questionnaire that could be answered meaningfully and anonymously by the respondents. A random sample of 20% of the managers was then taken. Managers were operationally defined as those who direct the work of other supervisors and/or other managers. Questionnaires were sent to 440 managers. Only 330 managers returned the questionnaires. However, only 304 questionnaires were usable. Hence, the usable response rate was 69%. Of the 304 respondents, 181 reported only to one superior while 123 reported to two or more superiors. While this fact was determined solely by means of the questionnaire, the analysis of the organizational chart indicated that each manager should be reporting only to one superior. Hence, violations of the principle of unity of command seemed to exist, at least as measured by the responses to this issue by the managers involved in the present study.

---

[1] The Civil Service Commission provided support and assistance for this research in a government agency.

[2] The authors wish to thank Rodney Morton for processing the data.

Requests for reprints should be sent to Martin J. Gannon, College of Business and Public Administration, University of Maryland, College Park, Maryland 20742.

[3] Both authors are also at the Bureau of Personnel Management Evaluation, Civil Service Commission, Washington, D.C.

The questionnaire was designed to measure several characteristics of managerial and organizational effectiveness which have been identified as midrange criteria (see House & Rizzo, 1972; Mahoney & Weitzel, 1969; Paine, 1970). The questionnaire included items touching on the dimensions of organizational effectiveness such as employee utilization, planning, selection based on ability, and role conflict and ambiguity. Many of the items were taken from the validated organizational practices' scales developed by House and Rizzo and by Mahoney and Weitzel. However, some of the items were selected because of the unique characteristics and problems facing the organization under study.

The questionnaire contained 50 attitudinal items. Subjects could respond to each item in terms of a 5-point scale of agreement (1) to disagreement (5). These items were then factor analyzed by means of the principal-component method, orthogonal rotation (BMD 03M). The eigenvalue specified in this rotation was 1.0; the commonality estimate for the diagonal was 1. Eleven factors were derived. Two criteria were then employed in order to interpret the derived factors. First, only items loading greater than .35 were considered. Second, complex items—those with relatively high loadings on two or more factors—were not used for interpretation. The *t* test for independent samples was then used to test the significance of the difference between the mean factor (standard) scores for each factor for those managers reporting only to one superior and for those managers reporting to more than one superior.

RESULTS[4]

Eleven rotated factors accounted for 62% of the variance. Factor 1 seemed to be concerned with the concept of selection based on ability and includes 9 of the 50 items. The items with the heaviest loadings were: (*a*) Higher level jobs are available for people who are capable (.75). (*b*) People who do a good job are rewarded (.72). (*c*) Promotions are based on ability and talent (.70).

Initiative-flexibility appeared to be reflective of Factor 2. The items with the heaviest loadings were: (*a*) People ask for assignments on their own (.73). (*b*) People accept changes in instructions readily (.40).

Factor 3 seemed to be indicative of role conflict and pressure. The items loaded heavily on this factor were: (*a*) People have questions about who is really their supervisor (.66). (*b*) People must be pressured to do the job right (.64). (*c*) People get conflicting assignments from higher up (.46). Role ambiguity or clarity underlied Factor 4, as exemplified by such items as: (*a*) People have a pretty clear idea of the quality expected in their work (.76). (*b*) People know pretty clearly what is expected of them (.60).

Employee utilization seemed to be the basis for Factor 5, as two items were loaded heavily on it: (*a*) People feel their jobs are too narrow and specialized (−.74). (*b*) People do too much unnecessary work (−.46). Factor 6 was indicative of coordination. The items loaded heavily on it were: (*a*) People are not receiving the help they need to learn the job right (−.69). (*b*) People can't seem to agree on just what they are trying to accomplish (−.58). (*c*) Work that is started is often never completed (−.54).

Factor 7 was loaded heavily by one item: People are not satisfied with performance appraisals (.84). Factor 8 was illustrative of adequacy of delegation. The items with heavy loadings included: (*a*) Higher level people do too much lower level work that could be given to others (−.71). (*b*) People must get approval for decisions that they should be able to make themselves (−.50). Factor 9 included two items with heavy loadings that are concerned with employee effort: (*a*) People put in more effort than they are required to do (.77). (*b*) People do their jobs without too much follow-up and checking

---

[4] More detailed statistical tables and a comparison of the factors derived in this study to those of the Mahoney and Weitzel (1969) study and to those of the House and Rizzo (1970) study may be obtained from the first author.

(.64). Factor 10 focused on planning activity. The items loaded heavily on this factor were: (*a*) Jobs are planned before they are started (.64). (*b*) People are asked for suggestions on how to improve the work (.60). Factor 11 seemed to involve employee development. The items with heavy loadings were: (*a*) There are not enough people who can do good work (−.62). (*b*) People do a quality job (.52).

Results of the *t* test indicated that four of the 11 factors derived in the analysis were statistically significant, and all of these relationships were in the predicted directions. Thus, managers with two or more superiors were significantly different from managers with only one superior in that they tended to feel: (*a*) selections are not based on ability ($t = 2.27$, $p < .01$). (*b*) Role conflict and pressure are high ($t = 2.54$, $p < .01$). (*c*) There is need for coordination ($t = 3.03$, $p < .01$). (*d*) More delegation is needed, that is, there should be more responsibility and autonomy associated with each job ($t = 2.70$, $p < .01$).

DISCUSSION

The present study has added support to the proposition that the strict use of the principle of unity of command is beneficial if the technology demands a bureaucratic structure. In the government agency under study, the violation of the principle of unity of command was related to dysfunctional consequences as measured by the attitudinal orientations of managers.

In summary, it appears that a mediating technology which links clients or customers together in an interdependent manner demands the strict implementation of the principle of unity of command. In such a situation it appears that subordinates want to report only to one superior, since the work requirements are such as to demand a well-defined bureaucratic structure. While other technologies may necessitate different strategies, it seems that a mediating technology can achieve organizational efficiency and effectiveness if subordinates operate in a situation where they can clearly identify their superior.

REFERENCES

FILLEY, A., & HOUSE, R. *Managerial process and organizational behavior.* Glenview, Ill.: Scott, Foresman, 1969.

HOUSE, R., & RIZZO, J. Toward the measurement of organizational practices: Scale development and validation. *Journal of Applied Psychology,* 1972, **56**, 388–396.

KAHN, R., WOLFE, D., QUINN, R., & SNOEK, J. *Organizational stress.* New York: Wiley, 1964.

LAWRENCE, P., & LORSCH, J. *Organization and environment.* Cambridge, Mass.: Harvard Graduate School of Business Administration, 1967.

MAHONEY, T., & WEITZEL, W. Managerial models of organizational effectiveness. *Administrative Science Quarterly,* 1969, **14**, 357–365.

PAINE, F. Organizational assessment dimensions as related to overall effectiveness of assessment teams. In M. Frey (Ed.), *Proceedings: Management research and practice.* Amherst, Mass.: Eastern Academy of Management, 1970.

RIZZO, J., HOUSE, R., & LIRTZMAN, S. Role conflict and ambiguity. *Administrative Science Quarterly,* 1970, **15**, 150–163.

THOMPSON, J. *Organizations in action.* New York: McGraw-Hill, 1967.

WOODWARD, J. *Industrial organization.* New York: Oxford University Press, 1965.

(Received March 12, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 395–397

REPETITION OF HIGHLY SIMILAR MESSAGES AND ATTITUDE CHANGE[1]

LEE McCULLOUGH[2]
Hay Associates, Philadelphia, Pennsylvania

THOMAS M. OSTROM
Ohio State University

Message repetition was predicted to increase positive attitudes in a situation where highly similar communications are used. Five similar advertisements (i.e., those using the same basic arguments but differing in the phrasing and order of points raised) served as stimulus messages. These ads were sequentially presented with attitude being measured by a cognitive response analysis of thoughts recorded by subjects attending to each message. Results supported the prediction of a positive relationship between the number of presentations and attitude.

Advertisers have long advocated repeated exposure of the same ad and attempted to develop measures to investigate the efficacy of this approach (Krugman, 1968). Several laboratory studies (Goldberg (1954), Wilson and Miller (1968), Johnson and Watkins (1971), however, have shown that repeated presentation of a persuasive message does not produce more immediate attitude change than does a single presentation. An important similarity among these failures to obtain a repetition effect is that they all repeated identical communications, and consequently, the subjects did not receive new information or arguments as a function of repetition. Sears and Freedman (1965) found evidence that people are more willing to change their attitudes when they think a message contains new information than when they expect a message to repeat previously encountered information. It may be, then, that the repetition effect can be obtained when similar but nonidentical messages are used.

There are two reasons for believing that the repetition effect might emerge for similar messages. First, Sears and Freedman (1965) hypothesized that the expectation of new information provides a satisfactory justification for relinquishing previous commitments, thereby allowing greater agreement with the advocated position. Second, people attend more closely to the message. For example, Brock, Albert, and Becker (1970) demonstrated that subjects show higher selective attention to a new communication advocating a known stand than to a previously encountered message, even when holding such factors as message utility and supportiveness constant. Thus, it was hypothesized that in a situation in which highly persuasive communications were used as stimuli, a positive relationship is found between the number of message repetitions and increasingly favorable subject attitudes.

METHOD

Subjects

The subjects were 57 introductory psychology student volunteers from Ohio State University, who viewed the ads while in groups of from one to five persons. Data from seven subjects were discarded before analysis because of their failure to correctly follow instructions.

Communications

The persuasive communications were 10 magazine ads published between 1960 and 1965. Half of these ads argued for the purchase of the product, Yardley After Shave, while the other five solicited contributions to the United Service Organizations (USO). Within each unit the particular ads chosen used the same general themes to support an identical position. The five Yardley ads each mentioned that (*a*) the product was not a typical after-shave; (*b*) it contained agents to prevent infections; (*c*) the fragrance was pleasant; and (*d*) Yardley replaced skin moisture. The USO messages stressed the following four themes: (*a*) Loneliness of young men away from home, (*b*) their tough working conditions, (*c*) the large number of Americans in this position, and (*d*) a civilian's obligation to help.

---

[1] This research was supported by an Ohio State University Dissertation Year Fellowship to L. McCullough and an N.S.F. Grant GN-534 to T. Ostrom. The research was conducted in connection with a dissertation completed by the first author at Ohio State University.

[2] The authors wish to thank A. Greenwald, D. Gregorski, M. Hakel, and H. Pepinsky for their assistance. Computer time was provided by the Ohio State University Instruction and Research Computer Center.

Requests for reprints should be sent to J. Lee McCullough, Hay Associates, 1845 Walnut St., Philadelphia, Pennsylvania 19103.

The various argument topics within a unit, though present in all advertisements, were phrased in different ways and sometimes appeared in different orders. Each of the messages was also dominated by a different photograph or design related to the attitude object.

Procedure

Subjects were informed that they would see five ads, each of which had been broken into four parts and were instructed to write every thought, comment, or opinion they had about an ad on a separate line in a standardized response booklet. A new response booklet, prefaced by a section repeating these directions and stressing that proper grammar and punctuation were not important, was provided for each advertisement.

Two groups of 25 subjects viewed either the five Yardley or USO messages. For each advertisement, the subjects first very briefly saw an overall picture of the message. After this brief orientation, the experimenter then incrementally exposed the ad through a series of four slides. The first of these slides, which was projected for 45 seconds, showed the graphic dominating the advertisement and the first segment of the written message. While this slide was being shown, the subjects recorded their reactions in the response booklet. At the end of the 45-second period the subjects were told that they could continue writing until they were finished recording their reactions. The second slide showing the same material as the first, plus the next segment of the message, was then projected for 45 seconds. The beginning of the new message segment was indicated by a small triangle. The subjects again recorded their reactions, after which the last two parts of the ad were exposed using the same procedure. Following the first ad the experimenter told the subjects that each message was to be considered separately and for subsequent ads they were not to refer to previous communications. The four remaining Yardley or USO messages were then presented.

Five different sequences of the five ads within a unit were prepared according to Latin square counterbalancing. Five subjects viewed each sequence. The four parts into which each ad was divided were planned to basically correspond to the four general themes, previously noted, that were present in every ad.

Dependent Variable

The dependent variable was the subject's net cognitive response score for each message (cf. Greenwald, 1968). To obtain this score, a rater determined the number of favorable, unfavorable and neutral responses recorded by a subject and computed the net value by subtracting the number of negative from the number of positive responses. A single response was defined as expressing a complete thought, fact, value, good or bad feature, feeling, or belief about the position advanced by the advertisement or about the ad itself. Cullen (1968) has found that a cognitive response measure significantly correlates with other "traditional" attitude indicators across a variety of issues.

The scoring technique was found to be reliable. Four days after the initial scoring the rater rescored 10 randomly selected response booklets from each of the two units. The test-retest correlations were .97 and .99 for the USO and Yardley units, respectively. Ten other randomly selected response booklets were scored by a second rater, resulting in interrater correlations of .83 (USO) and .90 (Yardley).

RESULTS

Net cognitive response scores were subjected to a three-way analysis of variance containing the within-subjects effect of repetition (number of similar ads presented) and the between-subject factors of ad type (either Yardley or USO) and counterbalanced order (within each unit ads were presented in five orders). Reference to Table 1 indicates that, as predicted, the overall mean cognitive response score was positively related to the number of previously viewed ads (linear trend $F = 4.55$, $df = 1/160$, $p < .05$). In addition USO ads were found to receive somewhat more favorable evaluations than did the Yardley messages ($F = 3.47$, $df = 1/40$, $p < .10$). None of the other compo-

TABLE 1

MEAN ATTITUDE AS A FUNCTION OF REPETITION AND MESSAGE TOPIC

Topic	Repetition				
	1	2	3	4	5
Yardley	−3.12	−2.44	−1.24	−.84	−1.88
USO	− .64	.24	.28	.92	1.40
Overall	−1.88	−1.10	− .48	.04	− .24

Note. USO = United Service Organizations.

nents, including the Ad Type × Repetition interaction, was significant.

It could be argued that the score obtained by subtracting negative from positive cognitive responses might fail to differentiate between a situation in which subjects actually became more positive and one in which they simply attenuated their negative responses as the experiment progressed. To examine such a possibility, positive and negative responses were tallied separately and then averaged over both topics. Table 2 indicates that the increasing favorability of net scores resulted not just from a restriction of negative reactions but rather from a tendency for positive responses to increase while negative responses decreased. In addition Table 2 shows that the total number of scored responses remained relatively constant over repetitions, thus suggesting that there was no decrement in attention to later ads.

DISCUSSION

The present results show that when similar rather than identical messages were employed, repetition did have an immediate effect on postexposure attitudes. Although this finding needs to be replicated with subjects drawn from other populations, several implications deserve mention. When a fixed number of repetitions are planned, it appears preferable to maximize the number of different advertisements that are included in the series. Also, the different ads need not actually contain different arguments. If the crucial factor is simply the anticipation of new information (to reduce commitment and facilitate attention), then one needs only to present the old arguments in a new context and with slightly altered phrasing.

TABLE 2

FREQUENCY OF POSITIVE AND NEGATIVE RESPONSES AS A FUNCTION OF REPETITION

Response	Repetition				
	1	2	3	4	5
Positive	2.76	3.24	3.22	3.58	3.70
Negative	4.64	4.34	3.70	3.54	3.94
Total	7.40	7.58	6.92	7.12	7.64

REFERENCES

BROCK, T. C., ALBERT, S. M., & BECKER, L. A. Familiarity, utility, and supportiveness as determinants of information receptivity. *Journal of Personality and Social Psychology,* 1970, **14**, 292–301.

CULLEN, D. A. Attitude measurement by cognitive sampling. Unpublished doctoral dissertation, Ohio State University, 1968.

GOLDBERG, S. C. Three situational determinants of conformity to social norms. *Journal of Abnormal Social Psychology,* 1954, **49**, 325–329.

GREENWALD, A. G. Cognitive learning, cognitive response to persuasion, and attitude change. In A. Greenwald, T. Brock, & T. Ostrom (Eds.), *Psychological foundations of attitude.* New York: Academic Press, 1968.

JOHNSON, H. H., & WATKINS, T. A. The effects of message repetitions on immediate and delayed attitude change. *Psychonomic Science,* 1971, **22**, 101–103.

KRUGMAN, H. E. Proceses underlying exposure to advertising. *American Psychologist,* 1968, **23**, 245–253.

SEARS, D. W., & FREEDMAN, J. L. Effects of expected familiarity with arguments upon opinion change and selective exposure. *Journal of Personality and Social Psychology,* 1965, **2**, 420–426.

WILSON, W., & MILLER, H. Repetition, order of presentation, and timing of arguments and measures as determinants of opinion change. *Journal of Personality and Social Psychology,* 1968, **9**, 184–188.

(Received May 8, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 398–400

DEVELOPMENT OF SCALES TO MEASURE ATTITUDES TOWARD MARIJUANA AND MARIJUANA USERS

VICTORIA COLAIUTA AND GEORGE BREED [1]

University of South Dakota

Four hundred and fifty students at the University of South Dakota served as subjects in the two-phase development of a 22-item marijuana attitude scale. The final form of the scale, partitioned into three factorially-defined, mutually exclusive subscales, deals with (*a*) the psychological, physiological, and sociological effects of marijuana on users; (*b*) marijuana as a psychosocial panacea; (*c*) liberal-conservative ambivalence toward the usage of marijuana.

Much of the information about attitudes toward marijuana has derived from attitude scales incorporated into questionnaires (DeFleur & Garrett, 1970; Eells, 1968; Gallup, 1969; Popoff, 1970). However, little evidence exists concerning the reliability and validity of these scales. Consequently, generalizations based on such data are limited.

The purpose of this study was to develop a set of reliable subscales for measuring several dimensions of attitudes toward marijuana.

METHOD

Phase 1

Subjects. Three hundred and seventeen undergraduate and graduate students served as volunteer subjects.

Procedure. Marijuana attitude scales have heretofore contained items relevant to the psychopharmacological effects and/or legal status of that substance (Eells, 1968; Gallup, 1969; Popoff, 1970). Both types of statements were incorporated in the present study, but for the coverage of attitudes in these areas additional items were included. Some dealt with the roles of the federal, state, and local governments as monitors, enforcers, and prosecutors under present law. Other statements had to do with attitudes about marijuana as a first step toward the use of "hard" drugs, as a carcinogen, and as a causative agent in birth defects. To increase coverage of a third content area, psychosocial aspects of marijuana use, items about marijuana as a social problem, and about the possible effects of marijuana use on the social development and adjustment of individuals, were added.

Forty-one items were developed in this manner. To reduce the possibility of response bias, 18 of the statements were worded so that agreement indicated a positive attitude toward marijuana, and 23 statements were phrased so that agreement indicated a negative attitude. Twenty graduate psychology students served as judges in evaluating the 5-point Likert-type statements as favorable or unfavorable toward marijuana. Next, sets of 21 and 20 of the statements were randomly assigned to each of two different forms of the scale (Forms "F" and "S", respectively).

Form F was administered to 155 students, Form S to 162 students. The correlation matrix of each form of the scale was factor analyzed by the basic-structure-successive factor method with varimax rotation (Horst, 1965), using squared multiple correlation communality estimates (Harman, 1969). An adaptation of the goodness of fit test for the factor model (Morrison, 1967) was computed for eight successive factors extracted in each analysis.

Nonsignificant chi-square/degree of freedom values, an eigenvalue cutoff of .95, and the relative drops in the percentage of variance accounted for by successive factors were criteria that indicated the number of factors actually warranted for each form was less than the original eight extracted. Therefore, a second factor analysis was carried out in which four factors were extracted for Form F and three were extracted for Form S.

Each factor defined a subscale, with the absolute maximum factor loading of an item across factors determining the subscale in which it would be exclusively included. Item-total correlations with the effect of the item removed from the total score [2] ($r_{X(T-X)}$) were computed for each statement within a subscale (McNemar,

---

[1] The authors wish to express their appreciation to Douglas K. Spiegel for his assistance in the analysis of the data.

Requests for reprints should be sent to George Breed, Department of Psychology, University of South Dakota, Vermillion, South Dakota 57069.

[2] Total score refers to a total subscale score, not the total score for all items in all subscales.

1969). Several statements with $r_{X(T-X)}$ values equal to or less than .30 were omitted, while the 34 statements retained were combined into a single scale for Phase 2 of the study. Phase 2 was performed to derive a single form of the scale involving various subscales.

Phase 2

Subjects. The subjects were 137 male and female undergraduate and graduate student volunteers. No subject had taken part in Phase 1.

Procedure. Instructions given the subjects and the procedures adhered to in analyzing data were identical to those employed in the initial phase of the study.

RESULTS AND DISCUSSION

Four factors were extracted and rotated in the final factor analysis of the attitude scale; however, any item with an $r_{X(T-X)}$ value of equal to or less than .30 was discarded. This resulted in the elimination of Factor 3. Twenty-two items comprised the final three subscales.

Subscales of the final attitude scale dealt with (*a*) the psychological, physiological, and sociological effects of marijuana on users (Effects) (eight items); (*b*) marijuana as a psychosocial panacea (Panacea) (six items); and (*c*) liberal-conservative ambivalence toward the usage of marijuana (Ambivalence) (eight items). Examples of the items included within each subscale with factor loadings are presented in Table 1.[3]

Items within subscale Effects have marijuana users as their attitude object. Agreement with the items indicates that the effects of marijuana on users are regarded as both personally and socially detrimental and that those individuals who use marijuana should be held legally responsible for their actions.

The statements comprising subscale Panacea neasured attitudes toward marijuana per se. In Panacea, marijuana is established as beneficial oth to the individual and to the society. Beause marijuana produces desirable effects and onsequences, usage is encouraged.

In the third subscale, Ambivalence, a central ttitude of personal conservatism is coupled with ne feeling that others should be allowed to use arijuana without the threat of legal penalty if ey so choose. In fact, each individual should e free to do "whatever he wishes" about marijuana. Ambivalence may indeed closely represent the attitudes of a number of people toward marijuana—specifically, that they are not sure how they feel about marijuana, given conflicting information about the positive and negative effects of the substance. Thus, they are cautious regarding personal use, but are willing to allow others to do as they please.

TABLE 1

EXAMPLES OF SUBSCALE STATEMENTS AND THEIR FACTOR LOADINGS

Statement	Factor loading
Factor I (subscale Effects)	
Marijuana has been found to cause birth defects.	.72
Anyone who uses marijuana is immature.	.66
Use of marijuana results in self-destruction.	.66
Factor II (subscale Panacea)	
Everyone should be encouraged to try marijuana.	.83
Marijuana brings out the best qualities in men.	.63
Marijuana is the single thing in the world today that can help promote love and brotherhood among nations.	.62
Factor IV (subscale Ambivalence)	
A person should be allowed to use marijuana without the threat of legal penalty if he so chooses.	−.69
Marijuana should be avoided at any cost.	−.69
The slight benefits of marijuana hardly justify its use.	−.65

The factors Effects, Panacea, and Ambivalence account for 34%, 11%, and 3% of the variance in the total scale, respectively. Estimates of internal consistency (coefficient alpha) for the respective subscales were .90, .86, and .92.

Potential users of the scale should keep in mind that the scale is based on the responses of South Dakota college students. Until other populations are sampled, the scale's general applicability should be viewed with caution. In constructing the present scale, efforts were made to include items from all relevant content areas (psychopharmacology, law, enforcement, physical and social effects of marijuana). Future development of the scale should include a further broadening of item content.

While the present analysis bears on the content validity of the scale, the degree of predictive validity and the ability to discriminate among various classifications of marijuana users and nonusers need to be determined.

---

[3] More detailed information on scale characterics can be obtained from the authors.

REFERENCES

DeFleur, L. B., & Garrett, G. P. Dimensions of marijuana use. *Journal of Counseling Psychology*, 1970, **17**, 468–476.

Eells, K. Marijuana and LSD: A survey of one college campus. *Journal of Counseling Psychology*, 1968, **15**, 459–467.

Gallup. Poll published in the *New York Times*, October 23, 1969.

Harman, H. H. *Modern factor analysis*. (Rev. 2nd ed.) Chicago: University of Chicago Press, 1969.

Horst, P. *Factor analysis of data matrics*. New York: Holt, Rinehart and Winston, 1965.

McNemar, Q. *Psychological statistics*. (4th ed.) New York: Wiley, 1969.

Morrison, D. F. *Multivariate statistical methods*. New York: McGraw-Hill, 1967.

Popoff, D. Feedback on drugs. *Psychology Today*, 1970, **3**, 51–52.

(Received March 22, 1973)

ANNOUNCEMENT

As of June 1, 1974 the editorial office of the *Journal of Applied Psychology* is moving to a new address. Please send all manuscripts and correspondence regarding manuscripts to Dr. Edwin A. Fleishman, Editor, American Institutes for Research, 3301 New Mexico Avenue, N.W., Washington, D.C. 20016.

Journal of Applied Psychology
1974, Vol. 59, No. 3, 401–403

LANDSCAPE OFFICE ATTITUDES:
REFLECTIONS OF PERCEIVED DEGREE OF CONTROL OVER TRANSACTIONS WITH THE ENVIRONMENT

MICHAEL W. McCARREY,[1] LYLE PETERSON,[2] SHIRLEY EDWARDS,
AND PAUL VON KULMIZ

Personnel Psychology Centre
Public Service Commission of Canada, Ottawa, Ontario

The component structure of attitudes toward a new office development concept, open office landscaping, were examined. A principal-components factor analysis of attitude data from 600 respondents identified the following six first-order oblique factors: (*a*) greater job satisfaction, (*b*) more effective communication, (*c*) greater productivity, (*d*) lack of auditory privacy, (*e*) lack of personal privacy, and (*f*) lack of confidentiality. An underlying bipolar second-order component termed "degree of perceived control over one's transactions with the environment" was found to underlie the attitudes.

A major trend in the design of office accommodation has been the use of open office landscape format with acoustically screened off work spaces replacing conventional walled offices. Its controversial nature is typified by conflicting research reports of its positive and negative consequences (e.g., greater communication and cohesion versus greater noise and loss of privacy; Cereghino, 1970; Riland & Kurtz, 1968).

This study represents an attempt to provide a preliminary theoretical structure which underlies and relates this conflicting feedback. It is hypothesized that a single bipolar theoretical factor, namely, the degree of control a person perceives himself to have over his two-way transactions with his environment, underlies the often cited positive and negative office landscape evaluation feedback.

METHOD

Participants

The participants were 600 employees representative of the total 1,130 (except for an underrepresentatation of support staff) of a Canadian Federal Government Department housed in a modern office tower centrally located in a large metropolitan area. Office planning and system analysis expertise had been employed in the development of the landscape office design and implementation, and, overall, 120 square feet of floor area had been allocated per employee.

Measuring Instrument

A 53-item Likert format, Open Office Opinion Inventory, was developed on the basis of (*a*) earlier measuring instruments (Allison, 1970; Riland & Kurtz, 1968), and (*b*) departmental input from representatives of each branch. Content included items designed to assess attitudes toward the visual layout, the noise level, the communication facility, perceived solidarity, job satisfaction, clarity of work space definition, and so on.

The reliability of inventory was checked by examining both internal consistency and test–retest data. In the case of internal consistency, the response data of the 600 subjects on seven randomly chosen repeated items (slightly different format) yielded a contingency coefficient of $C = .72$ with a chi-square test ($p < .001$). The response data of 28 subjects who took the complete inventory twice yielded a test–retest contingency coefficient of $C = .67$ for all 53 items over a 24-day time interval. The response data thus appeared to possess acceptable consistency and stability over time.

The validity of the response data was checked by first looking at the extent to which acquiescence and social desirability response sets were operative. Contingency coefficient estimates of the relationship between item endorsement and item bias and social desirability cue value were $C = .13$ and $C = .24$, respectively. The construct validity of the 53-Likert items was investigated by intercorrelating (point-biserial r) the 600 responses generating a 53 × 53 matrix. A principal-components analysis (little Jiffy Mark IV, Kaiser, 1970) of this matrix yielded empirical evidence that items intended to tap a specific characteristic did in fact cluster with other similar items, thus giving evidence that they had functioned in accord with the design intent.[3] The overall measure of sampling adequacy was .90, while the index of factorial simplicity was .78.

---

[1] Requests for reprints should be sent to Michael W. McCarrey, who is now at, Faculty of Psychology, University of Ottawa, Ottawa 8, Ontario, Canada.

[2] Now at the University of Ottawa, Ontario, Canada.

RESULTS AND DISCUSSION

The exploratory principal-components analysis of Kaiser was run on the 53 × 53 matrix, and orthoblique rotation was performed. Six main components were obtained and rotated (see Footnote 3).

Component 1 appears an overall preference factor combined with greater job satisfaction, while Component 2 seems the familiar more effective communication, solidarity and contract feature. Component 3 seems to indicate the recurrent greater productivity variable in this feedback data, while Component 4 appears the often repeated lack of auditory privacy feature. The recurring lack of personal privacy consequence of landscape office appears to have come out as Component 5, while lack of confidentiality communications content seems indicated by Component 6.

The component intercorrelation matrix showed all first-order components to be significantly related to each other (see Footnote 3). A second principal-components analysis (Little Jiffy Mark IV) was run on the 6 × 6 first-order matrix. Only one eigenvalue was greater than unity so the analysis identified a single second-order component that accounted for 67% of the variance.

It is suggested that this single second-order component reflects an underlying perceived degree of control over the individual's two-way transactions with others in the office landscape environment. Conceptually it appears the item definers for the six first-order components provide evidence to support the idea that these six are attitudinal reflections of one underlying construct, namely, degree of perceived control over transactions with the environment.

Looking at the negative pole, it is suggested that when respondents in landscape office perceived themselves to have inadequate control over input to and from the environment, they reported (*a*) lack of auditory privacy (second-order factor loading, −.77), (*b*) lack of personal privacy (factor loading, −.71), and (*c*) lack of confidentiality of communications content (factor loading, −.58) as consequences.

In other words, too much noise, too many distractions, lack of ability to put sustained effort on a problem, lack of territory definition, overly great accessibility to others, and so on, appear empirical reflections of a perceived inadequate degree of control over their own and others' transactions with the environment. This appears thus evidence of an "information overload phenomenon" that seems the consequence of continual exposure to unpredictable stimulation (noise) from a landscape environment which may continually alert and irritate the individual in his work space. In the event that individuals cannot take appropriate compensatory steps to deal with the greater constraints put on their behavior, it appears they report the familiar loss of visual, auditory and personal privacy.

On the other hand let us look at the positive pole and examine the attitudinal feedback from office landscape incumbents who may have suffered no such loss of control over interaction with their environment. It is hypothesized that the item definers for Components 1, 2, and 3 (second-order factor loadings of .90, .67, .93, respectively), such as, permits greater solidarity, offers greater contact, permits greater job satisfaction, am more productive, overlie an adequate level of perceived control over two-way transactions with the environment.

How can the same office landscape environment induce *both* the hypothesized adequate and inadequate perceived control over environmental transactions in it? Barker (1968) suggested the same environmental unit provides differential input to different individuals in it. These differences may be a function of the individual's behavior in the environment. By way of example, the environmental "press" or inputs on a motorist traveling at 95 miles per hour is not the same as that of another traveling at 60 miles per hour even when both are in the same traffic system on the same expressway. This research appears a rich area for further exploration both empirical and theoretical.

In retrospect then, it has been suggested the positive and negative attitudinal feedback data concerning an office landscape experience of 600 respondents is reflective of an underlying more basic perceived adequacy of control over transactions with the environment. One second-order component underlies Components 1 to 6, such that greater job satisfaction, more effective communications and greater productivity reflect an adequate ability to control interaction with one's environment. However, lack of visual and auditory privacy, lack of confidentiality of communications, lack of territory definition and lack of freedom interruptions reflect an inadequate ability to control interaction with one's environment, resulting in a possible "information overload phenomena." A consequence of such adequate or inadequate control over one's environment is the perception of greater or poorer job performance that may well relate to an objective measurement of job performance.

---

[3] Tabular data is available from the authors on request.

REFERENCES

ALLISON, W. L. Employee reactions to landscape environment. (RAID Project) Unpublished manuscript, Kingston, Ontario: Unemployment Insurance Commission, 1970.

BARKER, R. G. *Ecological psychology*. Stanford, Calif.: Stanford University Press, 1968.

CEREGHINO, E. E. Wayerhaeuser's new offices. *The Office*, 1973, 75(6), 12–18; 120–121.

KAISER, H. F. A second generation little Jiffy. *Psychometrika*, 1970, **35**, 401–415.

RILAND, L. H., & KURTZ, J. Employee reactions to a landscape office environment. Unpublished monograph. Rochester, New York: Eastman Kodak, Industrial Relations Department, 1968.

(Received May 14, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 404-405

DEVELOPMENT OF A BIOGRAPHICAL INVENTORY BLANK TO PREDICT FAKING ON PERSONALITY TESTS[1]

JEFFREY COHEN[2] AND JOEL LEFKOWITZ[3]

Baruch College, City University of New York

Fourteen Biographical Inventory Blank (BIB) items significantly predicted the propensity to fake personality inventories in a socially desirable manner, as measured by the K Scale of the Minnesota Multiphasic Personal Inventory (MMPI). Item analyses were performed on the responses of 76 job applicants, and the derived scoring weights were cross-validated on the responses of 42 other job applicants ($r = .66$, $p < .01$).

The use of psychological tests for personnel selection is often criticized on ethical as well as empirical grounds. Personality tests are additionally criticized because they are subject to faking. Indeed, one writer indicates that faking constitutes the major source of systematic variance of most personality inventories (McGee, 1962).

A biographical inventory blank (BIB) which successfully predicts faking could serve four purposes: (*a*) a moderator of personality test validity (hence interpretation) for individuals, (*b*) a suppressor of invalid personality test variance, (*c*) a relatively unobtrusive independent personality (predictor) measure, and (*d*) a means of identifying those personality characteristics which reliably differentiate falsifiers from nonfalsifiers in any given situational context.

METHOD

In this study "faking" is defined as a special case of the social-desirability-response set, operationally measured by the K Scale of the Minnesota Multiphasic Personality Inventory (MMPI). The K Scale was arbitrarily chosen over other possible psychometric measures because of its widespread usage and acceptance in clinical practice, as opposed to instruments usually deemed suitable for only research purposes. Moreover, a "one shot" measure was purposely chosen, rather than operationalizing faking as a change in response following an induced set to fake. This was done because (*a*) the research was undertaken with job applicants in the real-life (i.e., nonresearch) process of applying for specific positions and (*b*) the investigators were interested in approximating a measure of actual faking rather than "fakability." A shortened MMPI was used, consisting of all 30 K Scale items and the 45 items immediately preceding and following those in the complete MMPI, for a total of 75 items.

The equivalence of the K Scale in the shortened form to that in the complete MMPI was estimated by administration of both measures to 20 undergraduate and graduate students. The order of presentation was reversed for half of the group. The time interval between administration was three to 10 weeks. The correlation between the two measures was .90.

The BIB consisted of 80 items derived from Division 14 of the American Psychological Association catalog of life history items (Glennon, Albright, & Owens, 1966).

Both measures were administered to 118 applicants for a variety of clerical, supervisory, and sales positions that averaged $8,000 annual salary. The measures were included as part of a larger battery administered by a consulting firm to which the applicants had been sent by their prospective employers.

Item analysis for the BIB was performed on a randomly selected group approximately twice as large as the hold-out group to be used for cross-validation ($n = 76$ and 42, respectively). The criterion group was dichotomized at the median K-scale score (18). Chi-square analyses, the vertical percent method, and Strong's tables of net weights converted to unit directional weights were used to develop the BIB (Guion, 1965).

RESULTS AND DISCUSSION

Fourteen BIB items were statistically significant (eight items, $p < .05$; six items, $p < .10$). Applying the derived weights for these items to the responses of the hold-out group yielded a correlation between the BIB and K-scale scores of .66 ($p < .01$).

---

[1] This study is based on the first author's master's thesis, written with the supervision of the second author. Also at BFS Psychological Associates, Inc.

[2] Now at Personnel Research and Development Corporation.

[3] Requests for reprints should be sent to Joel Lefkowitz, Psychology Department, Baruch College, City University of New York, 17 Lexington Avenue, New York, New York 10010.

The BIB profile of the predicted dissimulator is of one who is over 26 years of age, married, "occasionally" tells jokes, considers himself "middle of the road" politically, and is not bothered by being interrupted in the middle of a problem. He tends to ignore co-workers' bad habits and takes direct action when overcharged or not repaid on a loan. During his school years, his parents permitted most desires if obtained on his own, and called a physician when he was ill; he mastered difficult problems by independent reading; his most outstanding negative experience was lack of achievement in sports; and he went out one or fewer nights per week while in high school. His father was a businessman, and his wife is satisfied with their present living quarters.

These demographic characteristics can be taken at face value, that is, as personal attributes which reflect the developmental conditions that have fostered a propensity to fake personality inventory responses in a socially desirable manner. However, these items themselves may, in part, reflect primarily that same social-desirability-response set measured by the K Scale. (Ten of the 14 significant items request opinions or unverifiable facts which could be biased in a stereo-typically-socially-desirable direction. Nine of those 10 positively weighted alternatives are responses which imply independency, achievement orientation, and conventionality.) In order to determine whether such is the case, future research ought to include: (*a*) other potentially fakable BIB items (which, however, might *not* be predictive), (*b*) items which seek to ascertain the same information but in a more unobtrusive manner, and (*c*) items which are subject to external verification.

REFERENCES

GLENNON, J. R., ALBRIGHT, L. E., & OWENS, W. A. A catalog of life history items. Greensboro, N.C.: The Richardson Foundation, 1966.

GUION, R. M. *Personnel testing.* New York: McGraw-Hill, 1965.

MCGEE, R. L. The relationship between response style and personality variables. *Journal of Abnormal and Social Psychology,* 1962, **64**, 229–233.

(Received May 9, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 406-407

THE INFLUENCE OF TECHNICAL KNOWLEDGE ON POSITION ERROR IN RANKING

EDWIN E. WAGNER[1] AND THOMAS O. HOOVER

University of Akron

The biasing effect of position error in ranking was demonstrated in a situation approximating real-life conditions. It was also shown that position error is less likely to occur when subjects undertaking the evaluation are knowledgeable about the dimension being ranked.

Symonds (1936), who studied the influence of serial position on ranking error within the confines of an interest questionnaire, concluded that overranking tends to occur at the beginning of a ranking series and then progressively decreases. Recently, with an experimental design which more adequately controlled for context effect, it has been shown that there is a tendency to overevaluate the first and last positions when ranking a set of five entities. This terminal error occurred not only under theoretical conditions when subjects may have been relatively uninvolved with the stimuli to be ranked (Wagner & Hoover, 1974) but also under more meaningful circumstances which might have been expected to evoke strong personal feeling (Wagner & Hoover, 1974). Since ranking procedures are important for applied areas of psychology, further research on this phenomenon seemed warranted, particularly under conditions approximating real-life situations. The present study attempts to investigate the effect of technical knowledge (i.e., correct information relating to the entities and/or dimension to be ranked) on position error in a situation requiring the direct evaluation of live human beings.

METHOD

Five Air Force Reserved Officers' Training Corps (ROTC) cadets, members of the Saber Squadron, a military drill team at the University of Akron, were put through a six-minute drill sequence which was recorded and preserved for replay on videotape. The men were identified by affixing fictitious three-letter names to their backs (Tom, Dan, Jim, Ron, Bob), which were clearly discernible to the TV viewer. All military honors, rank, decorations and other means of identification were removed from their uniforms prior to the drill. The drill sequence was then presented via closed circuit TV to a control group of 120 general psychology undergraduates who had no previous exposure to service related experiences and were therefore considered to be relatively unskilled in judging military drill.

The experimental group was composed of 120 cadets in the Air Force or Army ROTC who had been trained in drill exercises and could therefore be considered expert observers. All subjects were instructed to rank the individual members of the drill team from one to five in terms of judged ability to properly execute the maneuvers. Evaluation sheets were then handed out to each observer. Ostensibly, these forms were all alike, but actually, the five names to be ranked were completely counterbalanced so that every combination occurred just once for each group, that is, $5! = 120$. In this manner it was possible to neutralize any context effect and isolate the influence of position per se on ranking error.

RESULTS AND DISCUSSION

Both the skilled and unskilled judges significantly differentiated among the five individual members of the drill team (see Table 1). Although median selection values varied, both groups ranked Tom first and Bob last, indicating some general agreement between groups despite differences in technical knowledge. Of greater interest, however, was the fact that significant

TABLE 1

MEDIAN AND FRIEDMAN TWO-WAY ANALYSES OF VARIANCE COMPARING THE RANKINGS OF FIVE CADETS BY UNSKILLED AND SKILLED JUDGES

Group	n	Cadets					df	χ^2_r
		Tom	Ron	Jim	Bob	Dan		
Unskilled	120	3.70	2.61	3.04	2.01	3.50	4	40.76*
Skilled	120	3.83	3.33	2.87	1.69	3.33	4	68.17*

* $p < .001$.

---

[1] Request for reprints should be sent to Edwin E. Wagner, Department of Psychology, University of Akron, Akron, Ohio 44325.

position differences occurred for the unskilled judges but not for the skilled judges, suggesting that position error is more likely to crop up when evaluators lack information relating to the ranking task (see Table 2). The practical implication of this finding seems clear: lack of expertise may not lead to random ranking error but rather to a bias which induces subjects to prefer one ranking position over another, thus accidentally favoring certain individuals for reasons unrelated to the purpose of the evaluation. In this experiment the bias seemed to be concentrated on the first ranking position which received a median value of 3.55. The observed error pattern confirms neither the declining linear model advocated by Symonds nor the terminal error model proposed by Wagner and Hoover; yet, it is apparent that position error has been demonstrated in a real-life situation, that is, the evaluation of live human beings on a meaningful task. Considering the extensive use of ranking procedures, particularly in industry and in the military, it seems that further efforts should be bent toward defining the parameters of position error and developing remedial and/or compensatory procedures.

TABLE 2

Medians and Friedman Two-Way Analyses of Variance Comparing the Rankings of Serial Position by Unskilled and Skilled Judges

Group	n	Serial position					df	χ^2_r
		1	2	3	4	5		
Unskilled	[illegible]	3.55	3.10	2.79	2.95	2.67	4	9.88*
Skilled	[illegible]	[illegible]	2.63	3.10	2.89	3.08	4	1.59

* $p < .05$.

REFERENCES

Symonds, P. M. Influence of order of presentation of items in ranking. *Journal of Educational Psychology*, 1936, **27**, 445–449.

Wagner, E. E., & Hoover, T. O. The effect of serial position on ranking error. *Educational and Psychological Measurement*, 1974, in press.

Wagner, E. E., & Hoover, T. O. Effect of terminal ranking error on meaningful occupational choice. *Journal of Applied Psychology*, 1974, **59**, 247–248.

(Received May 9, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 3, 408–410

BRIGHTNESS CONTRAST EFFECTS ON RECALL OF PROJECTED HIGHWAY SIGN-TYPE STIMULUS MATERIAL

ROBERT W. SWEZEY[1]

Applied Science Associates, Reston, Virginia

Stimulus-surround brightness contrast of projected highway sign-type stimuli was varied by manipulating photographic exposure time. Stimulus materials comprising six contrast levels of black on white (e.g., 97%, 92%, 90%, 85%, 67%, and 34%) were developed and reproduced on 35-millimeter slides. These materials were presented at two different stimulus-exposure-time conditions on each of two separate recall tasks. It was concluded that recall of projected materials is not seriously degraded as contrast decreases from 97% to 67%. Below the 67% contrast level, a significant degradation in recall was observed.

Effects of brightness contrast variations in character legibility have been widely investigated (Cornog & Rose, 1967; Tinker, 1963). Most studies have measured variables such as, visibility, apprehension, reading speed, and blink rate. These variables have historically been employed as criterion measures in legibility research (Tinker, 1963). Only a relatively small number of studies have employed recall and/or comprehension as legibility criteria. Research that includes these types of variables as criteria is generally lacking in the early literature. Recently, however a number of studies by Poulton (1959, 1965, 1969a, 1969b, 1972) and his associates have employed such variables as criteria. These studies have been concerned with variables such as printing types and formats (Poulton, 1959), letter differentiation (1965), type size (1969a, 1972), style and spacing (1972), etc. Poulton (1969b) studied brightness contrast effects on ability to skim lists of food ingredients.

The purpose of the present study was to investigate stimulus-surround brightness contrast on recall of projected highway sign-type stimuli. Stimulus-surround brightness contrast was varied across six levels on two separate tasks, and the effects of these variations on recall were assessed.

METHOD

Subjects

The subjects were 120 students enrolled in driver education courses at a suburban high school. Of these, 111 were in the junior class and nine were in the senior class. Sixty-six subjects were female and 54 were male. All subjects had visual acuity of 20/30 or better corrected, according to the Snellen chart.

Phases

The study was conducted in two phases. In Phase 1, five simulated highway sign stimuli were designed and reproduced on 35-millimeter slides at each of six contrast levels (e.g., approximately 97%, 92%, 90%, 85%, 67%, and 34%). In all cases the surround was white, and the stimuli were one of six levels of gray. Recall was measured according to a scoring system which allotted points for correctly recording directions, locations, and other information items on the slide. An independent group design was employed, in which each group of subjects viewed each projected stimulus for either six seconds or 12 seconds. The stimulus exposure was followed by a 36-second off time during which subjects noted on an answer form all information items that they could recall from the previous stimulus exposure. Phase 1 thus consisted of six groups of five subjects for each of the two exposure time conditions. Each group saw five stimulus slides all of which were at the same contrast level.

In Phase 2, one stimulus slide consisting of information on point assessments for driving violations in the State of Maryland was reproduced at each of the six contrast levels. Recall was measured by a 15-point multiple-choice test

---

[1] This study is based on a dissertation completed in partial fulfillment of the requirements for the doctoral degree at the University of Maryland, 1972. A more detailed description of the study is presented in that dissertation (Swezey, 1972).

The helpful comments of Irwin L. Goldstein, Nancy S. Anderson, C. J. Bartlett, and of John D. Folley, Jr. are gratefully acknowledged.

Requests for reprints should be addressed to R. W. Swezey, Applied Sciences Associates, Reston International Center, Reston, Virginia 22090.

constructed specifically for this purpose. Test–retest reliability of this measure was estimated at .88. The subjects were instructed to read through the projected stimulus one time *only* for one condition and to read for as long as was necessary for them to understand the contents for the second condition. In Phase 2, six independent groups of five subjects for each of the two exposure conditions viewed a single stimulus slide at one of the six contrast levels. Reading time in each exposure condition was measured via a stopwatch. The subjects' reading comprehension level was also measured using the Educational Testing Service Cooperative English Test on Reading Comprehension. This score served as the covariate in the covariance analysis experimental design.

The contrast levels were produced by manipulating photographic exposure time. All photographs were exposed at a lens setting of *f*/32 and developed for one minute. Six exposure times were employed, 4 seconds, 7 seconds, 8.5 seconds, 12 seconds, 17.5 seconds, and 45 seconds. This process effectively manipulated the brightness of the stimulus without affecting the brightness of the surround. Thus, six levels of stimulus-surround brightness contrast were obtained, and a series of photographic prints were manufactured. The prints were then copied onto 35-millimeter slide film. The resulting stimuli were then matched with Munsell percent reflectances, using a photovolt photometer. The scale used for matching was the Munsell neutral value 32-step matte scale. Once the brightness level of each stimulus was matched with a known Munsell value, the percentage of contrasts between stimulus and surround brightness were calculated using the Luckiesh (1937) formula. Letters ranged in size from approximately 3/16 inch to approximately 5/16 inch tall when projected. Distance from the eye to the midpoint of the stimulus display on the screen ranged from 76 inches to 83 inches, depending on seating position in the room.

RESULTS

For Phase 1 data (simulated highway sign recall) a 6 × 2 covariance analysis was calculated. Figure 1 shows these data. The independent variables were six contrast levels and two stimulus exposure times (six seconds and 12 seconds). The dependent variable was recall of the simulated highway signs (points correctly recalled out of a possible total of 80 points per stimulus slide). The covariate was Educational Testing Service Total Reading Comprehension score.

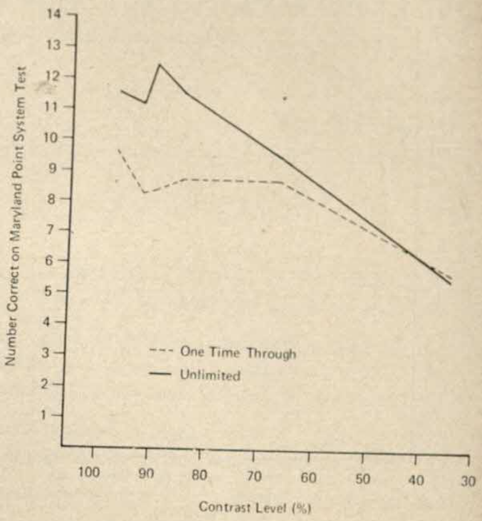


FIGURE 1. Highway sign phase results: Mean recall scores adjusted by covariate.

Results indicated a significant exposure time effect ($F = 7.70$, $df = 1/47$, $p < .01$) and a significant contrast effect ($F = 7.50$, $df = 5/47$, $p < .001$) with the Exposure Time × Contrast interaction being nonsignificant. The covariance analysis was conducted on mean recall scores across contrast and exposure time conditions. Scheffé's post hoc multiple-comparison tests were calculated on all two-way comparisons. These statistics indicated that the 34% contrast condition produced the least recall and differed significantly from the 67% condition ($p < .10$) and from the 85% condition ($p < .10$) for the six-second stimulus exposure. For the 12-second stimulus exposure, the 34% contrast condition differed significantly from the 67% condition ($p < .01$) and from the 90%, 92%, and 97% conditions ($p < .10$). No other two-way comparisons were found significant. Since the Scheffé test is an extremely conservative test, it was felt that the .10 level could be accepted as an appropriate indication of statistical significance (Edwards, 1960).

Omega-squared statistics indicated that 34.11% of the total variance could be accounted for by the contrast main effect and 7.3% of the total variance by the exposure time main effect. Trend analyses yielded significant linear and quadratic effects ($F = 22.95$, $p < .001$; $F = 15.12$, $p < .001$), respectively.

In Phase 2 (the Maryland Point System phase), a 6 × 2 covariance analysis indicated a significant stimulus exposure effect ($F = 9.88$, $df = 1/74$, $p < .01$), a significant contrast effect

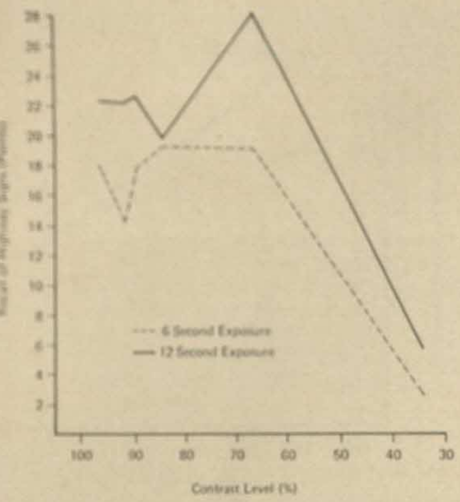


FIGURE 2. Point system phase results: Mean recall scores adjusted by covariate.

($F = 4.68$, $df = 5/47$, $p < .01$), and a significant Stimulus Exposure × Contrast interaction ($F = 8.42$, $df = 5/47$, $p < .001$). The independent variables were two stimulus exposure conditions (reading through the material one time only versus unlimited reading time) and six contrast levels; the dependent variable was the score on the Maryland Point System test; and the covariate was the Educational Testing Service Total Reading Comprehension score. Figure 2 shows these data.

Scheffé test results indicated no significant two-way comparisons for the one-time through stimulus exposure condition. For the unlimited time stimulus exposure condition, the 34% contrast condition again produced the least recall and differed significantly from the 90% condition ($p < .05$) and from the 85%, 92%, and 97% conditions ($p < .10$). No other two-way comparisons were significant. Omega-squared statistics indicated that 14.91% of the total variance could be accounted for by the contrast effect. Similarly 7.2% of the total variance was accounted for by the stimulus exposure effect, and 2.77% by the Stimulus Exposure × Contrast interaction. Trend analyses for the point system phase indicated a significant linear trend ($F = 24.71$, $p < .001$). The quadratic trend was nonsignificant.

DISCUSSION

Several summarizing concepts can be gained from the general contrast literature. First, it is fair to state that in the majority of achromatic *contrast direction* studies, black stimuli superimposed upon white backgrounds are generally found to be more legibile than the reverse (see Tinker, 1963). Second, a general conclusion has been that as brightness contrast between a stimulus and its surround decreases, the corresponding legibility also decreases in a monotonic fashion. Third, dark stimuli on light surrounds are to be preferred to light on dark regardless of hue and chroma. Fourth, in the case of black print on colored paper, minor variations in hue are of minimal importance as long as the contrast ratio remains high.

Results of the present study indicated that recall of projected alphanumeric material is not seriously degraded by stimuli which employ less than maximum stimulus-surround brightness contrast. For this aspect of legibility, high contrast materials produced no better recall than did considerably lower contrast materials. These results occurred over two types of recall tasks. The range of contrast conditions under which recall scores are not affected appears to extend from approximately 97% to approximately 67%.

REFERENCES

CORNOG, D. Y., & ROSE, E. C. Legibility of alphanumeric characters and other symbols: II. A reference handbook. (National Bureau of Standards Miscellaneous Publication No. 262-2) Washington, D.C.: U.S. Government Printing Office, 1967.

EDWARDS, A. L. *Experimental design in psychological research*. (Rev. ed.) New York: Holt, Rinehart & Winston, 1960.

LUCKIESH, M. The perfect reading page. *Electrical Engineering*, 1937, **56**, 779–781.

POULTON, E. C. Effects of printing types and formats on the comprehension of scientific journals. *Nature*, 1959, **184**, 1824–1825.

POULTON, E. C. Letter differentiation and rate of comprehension in reading. *Journal of Applied Psychology*, 1965, **49**, 358–362.

POULTON, E. C. Skimming lists of food ingredients in different sizes. *Journal of Applied Psychology*, 1969, **53**, 55–58. (a)

POULTON, E. C. Skimming lists of food ingredients printed in different brightness contrasts. *Journal of Applied Psychology*, 1969, **53**, 498–500. (b)

POULTON, E. C. Size, style, and vertical spacing in the legibility of small typefaces. *Journal of Applied Psychology*, 1972, **56**, 156–161.

SWEZEY, R. W. An investigation of stimulus-surround brightness contrast: Effects on recall of projected material. Unpublished doctoral dissertation, University of Maryland, 1972.

TINKER, M. A. *Legibility of print*. Ames: Iowa State University Press, 1963.

(Received May 10, 1973)

JOURNAL OF APPLIED PSYCHOLOGY

Copyright © 1974 by the American Psychological Association, Inc.

August 1974 Vol. 59, No. 4

ARTICLES

The Relationship Between Managerial Values and Managerial Success in the United States, Japan, India, and Australia *George W. England and Raymond Lee* 411

Influence of Dogmatism, Risk-taking Propensity, and Intelligence on Decision-making Strategies for a Sample of Industrial Managers *Ronald N. Taylor and Marvin D. Dunnette* 420

Quality of Decisions: Individuals versus Real and Synthetic Groups *Bertram Schoner, G. C. Hoyt, and Gerald L. Rose* 424

Brainstorming by Individuals, Coacting and Interacting Groups *Warren R. Street* 433

Cue Utilization and Expert Judgments: A Comparison of Independent Auditors with Other Judges *Robert H. Ashton* 437

Role of the Rater in Performance Appraisal *Richard J. Klimoski and Manuel London* 445

Realistic Expectations as an Aid in Reducing Voluntary Resignations *Daniel R. Ilgen and William Seely* 452

Probability and Utility Assumptions Underlying Use of the Strong Vocational Interest Blank *Frank L. Schmidt* 456

Consumer Dissatisfaction as a Function of Dissatisfaction with Interpersonal Relationships *Gerhard W. H. Scherf* 465

Seat Belts: Relations Between Beliefs, Attitude, and Use *Gunilla Fhanér and Monica Hane* 472

Effects of Caffeine on Alertness in Simulated Automobile Driving *Edmund G. Regina, Gene M. Smith, Charles G. Keiper, and Robert E. McKelvey* 483

Time Delay and Similarity Effects in Facial Recognition *Kenneth R. Laughery, Paul K. Fessler, David R. Lenorovitz, and Darryl A. Yoblick* 490

SHORT NOTES

Relationship of Role Conflict and Role Ambiguity to Job Involvement Measures *W. Clay Hamner and Henry L. Tosi* 497

Stress and Performance During and After Officer Candidate School *J. Richard Jennings, R. M. Rose, and Leo Kreuz* 500

A Performance Correlate of Personal Control Beliefs in an Organizational Context *W. J. Heisler* 504

Comment on Locke's Satisfiers and Dissatisfiers among White-Collar and Blue-Collar Employees *Robert G. Hadley* 507

Achievement Motivation among Managers of Small Businesses, Corporation Managers, and Business Students *John C. South* 509

Effects of Applicant's Sex and Difficulty of Job on Evaluations of Candidates for Managerial Positions *Benson Rosen and Thomas H. Jerdee* 511

Diary No 828
Date 18-11-74

(*Continued on next page*)

Spin-Offs from Behavioral Expectation Scale Procedures *Milton R. Bloo* 513

A Utility Analysis of a Weighted Application Blank Designed to Predict Turnover fo
Clerical Employees *Raymond Lee and Jerome M. Booth* 516

Prediction of Scuba Training Performance *Robert J. Biersner and David H. Ryma* 519

Predicting Alumni Financial Donation *Warren S. Blumenfeld and Patricia L. Sartai* 522

Effects of Spectral Differences in Illumination on Fatigue
James B. Maas, Jill K. Jayson, and Douglas A. Kleiber 524

Effect of Ambient Illumination on Noise Level of Groups
Mark Sanders, Jerry Gustanski, and Mike Lawton 527

Visual Dependency in the Erect and Supine Positions
Jacob H. Lichtenstein and Rayford T. Saucer 529

Effects of Numerical Data Upon Reading Time and Recall of Scientific Writing
William Wi 532

Type of Transmittal Letter and Questionnaire Color as Two Variables Influencing
Response Rates in a Mail Survey *Michael T. Matteson* 535

LIST OF MANUSCRIPTS ACCEPTED 52 528

SPECIAL ANNOUNCEMENT 515

ERRATUM 534

INSTRUCTIONS TO AUTHORS

Articles must possess: (*a*) conciseness; (*b*) unambiguous and simple vocabulary; (*c*) conformity to accepted technical style in tables, terminology, and references; and (*d*) conclusions that are clearly related to the evidence presented. Authors must progress from a statement of problem or purpose, through analysis of evidence, to conclusions and implications.

Criteria for evaluation include: (*a*) significance of contribution, (*b*) technical adequacy, (*c*) appropriateness for the *Journal of Applied Psychology*, and (*d*) clarity of presentation.

See the *Publication Manual of the American Psychological Association* (1967 Revision), $1.50, for the format of manuscripts. Articles not prepared in this manner cannot be reviewed. Note sections on typing (p. 48, *all copy must be double-spaced*), tables, figures, and references. The first author's last name must be typed on the upper left-hand corner of each page, except for blind reviews. Do not abbreviate, for example, journal titles in reference lists, the word "subject," names of groups, etc. Each manuscript copy must include an abstract of 100–120 words typed on a separate sheet of paper.

Optional blind review must be specifically requested when a manuscript is first submitted. The author's name and affiliation must appear only on a separate title page, included with each copy of the manuscript. Footnotes containing the identity of the author or his affiliation must be on separate pages.

Check recent issues of the *Journal of Applied Psychology* for approximate length of regular articles. (One printed page equals three double-spaced manuscript pages.) A few longer articles of special significance are printed as monographs. Short Notes feature brief reports on studies such as those involving some methodological contribution or important replication.

The original drawing of a figure or an 8 × 10 inch glossy print must be submitted with the original copy of the manuscript. Duplicate copies may be photographic or pencil drawn. Figures *must* be hand-lettered.

Supplementary materials formerly deposited with the National Auxiliary Publications Service need no longer be submitted with the manuscript. Retain supporting and raw data for at least 5 years after publication of the article.

Journal of Applied Psychology
1974, Vol. 59, No. , 411–419

THE RELATIONSHIP BETWEEN MANAGERIAL VALUES AND MANAGERIAL SUCCESS IN THE UNITED STATES, JAPAN, INDIA, AND AUSTRALIA

GEORGE W. ENGLAND[1] AND RAYMOND LEE
University of Minnesota

The relationships between managerial values and managerial success were investigated for a diverse sample of American managers ($n = 878$), Australian managers ($n = 301$), Indian managers ($n = 500$), and Japanese managers ($n = 312$). Cross-validated results showed that values patterns were significantly predictive of managerial success and could be used as a basis for selection and placement decisions. Results also indicated that managers from the four countries were rather similar in terms of the personal values that were related to success. More successful managers have pragmatic, dynamic, and achievement-oriented values, while less successful managers have more static and passive values. Arguments for and against using value assessment results in selection and placement decisions are discussed.

In recent years considerable effort has been expended in attempts to predict managerial effectiveness by using maximum performance measures and typical behavior measures. Most of this effort has focused on tests of (*a*) intellectual abilities, (*b*) spatial and mechanical aptitudes, (*c*) perceptual accuracy, (*d*) motor abilities, and (*e*) personality and interest. According to Ghiselli's (1966) review, each measure shows average *r*s in the range of .25 to .30. Another area of potential importance, but receiving less emphasis, has been the notion of predicting managerial success from managers' personal values. The significance of investigating the value systems of managers is seen when one considers the following assertions and their implications:

1. Personal value systems influence a manager's perception of situations and problems he faces.
2. Personal value systems influence a manager's decisions and solutions to problems.
3. Personal value systems influence the way in which a manager looks at other individuals and groups of individuals; thus they influence interpersonal relationships.
4. Personal value systems influence the perception of individual and organizational success as well as their achievement.
5. Personal value systems set the limits for the determination of what is and what is not ethical behavior by a manager.
6. Personal value systems influence the extent to which a manager accepts or resists organizational pressures and goals.
7. Some personal value systems may contribute to managerial performance, some may be irrelevant, and some may be antithetical to achievement efforts.

It seems reasonable, therefore, to expect that personal value systems will influence not only the behavior of managers but their success as managers. Despite the apparent logic of the value–performance relationship, little empirical research has related values and performance of managers. The present study investigates the relationship between values of managers and their success as managers in four countries. The major focus of the research is comparative in nature and centers on answering the following questions: (*a*) What is the nature of the relationship between managerial values and managerial success? (*b*) How similar or different are the relationships between values and success among managers in four different countries? (*c*) What are the distinguishing characteristics of value differences between successful managers and relatively less successful managers and what do these differences imply behaviorally? (*d*) What type of causal relationship is suggested? Are managers differentially successful because of their values or are their value systems different because they have achieved different levels of success?

---

[1] Requests for reprints should be sent to George W. England, Industrial Relations Center, 537 Business Administration Building, University of Minnesota, Minneapolis, Minnesota 55455.

METHOD

Samples

The data used in the present study were obtained by selecting national samples of managers in Australia, India, Japan, and the United States from major directories of corporation executives and directors. The national samples were developed to be roughly comparable and were selected on the basis of three stratifying variables: (*a*) size of organization in terms of number of employees, (*b*) level of the manager in the organization, and (*c*) organizational function of the manager. While detailed sampling plans and data collection procedures have been presented elsewhere: the United States (England, 1967), Japan (England & Koike 1970), India (England, Agarwal, & Dhingra, 1973), and Australia (England, 1972), it should be noted that the percentages of usable returns for the four countries were as follows: the United States (31%), Japan (26%), India (72%), and Australia (30%). Each national sample of managers represents a relatively diverse group in terms of organizational variables (type of company, size of firm, department or function, organizational level, and line-staff position) and in terms of personal variables (years of managerial experience, age, and formal education).

Value Measurement

A framework was developed by England (1967) to delineate the relationship of values to behavior for managers and was subsequently used in the development of a measurement approach to personal value systems. A personal value system is viewed as a relatively permanent perceptual framework which shapes and influences the general nature of an individual's behavior. Values are similar to attitudes but are more ingrained, permanent, and stable in nature; they are also more general and less tied to any specific referent than is the case with many attitudes.

Several major classes of overlapping values are recognized in the framework. All possible values that might be held by an individual or by a specific group constitute the total value space and are known as potential values. The potential values are made up of two classes of values: (*a*) nonrelevant or weak values for a specific group or individual (those values which would have little or no impact on behavior) and (*b*) conceived values (those which are likely to be translated from the intentional state into behavior). Conceived values are made up of (*a*) operative values (those which have a relatively high probability of being translated from the intentional state into actual behavior, (*b*) intended values (those which are viewed as important but may have only a moderate probability of being translated from the intentional state into behavior because of situational factors), and (*c*) adopted values (those which are less a part of the personality structure of the individual and affect behavior largely because of situational factors).

The development of the Personal Values Questionnaire was based on the rationale that the meanings attached by an individual to a carefully specified set of concepts provides a useful description of his personal value system, which may in turn be related to his behavior in systematic ways. This attempt to describe an individual's values through the use of a carefully specified set of concepts was influenced by the work of Osgood, Suci, and Tannenbaum (1957) and represents an adaptation of their methodology.

In order to specify a set of concepts relevant to the personal value systems of managers, a pool of 200 concepts was selected from literature dealing with organizations and with individual and group behavior. In addition, ideological and philosophical concepts were included to represent major belief systems. A panel of expert judges reduced this pool to a set of 96 concepts, which was further reduced to 66 concepts based on pilot studies. To provide a framework within which respondents could conveniently evaluate each concept, the concepts were categorized into five classes: goals of business organizations, personal goals of individuals, groups of people and institutions, ideas associated with people, and ideas about general topics.

In the Personal Values Questionnaire, four scales are used to represent two modes of valuation. Since the general value of an object or idea to an individual is thought to be largely a function of its degree of importance to him, the primary or power mode of valuation used is the importance scale which consists of three points—high, average, and low. Three secondary modes of valuation were also developed from the literature. The pragmatic mode of valuation suggests that an individual has an evaluative framework that is primarily guided by success–failure considerations: Will a certain course of action work or not? How successful or unsuccessful is it apt to be? The pragmatic mode of valuation runs throughout much of the literature dealing with managers and finds support in various analyses of the Allport-Vernon-Lindzey (1960), *Study of Values* which suggest a major dimension of values as being "pragmatic and utilitarian" or the finding of a major value factor being "idealism versus practicality."

The ethical–moral mode of valuation implies an evaluative framework consisting of ethical considerations influencing behavior toward actions and decisions that are judged to be "right" and away from those judged to be "wrong." The existence of a moralistic-ethical orientation is at the very heart of most religious belief systems and is supported by analyses of the *Study of Values* which find value dimensions or factors such as "social and altruistic" and "idealism."

The affect or feeling mode of valuation suggests an evaluative framework which is guided by hedonism; one behaves in ways that increase pleasure and decrease pain. The affective component of values has an extensive philosophical and psychological background and seemed important to include as an orientation in the present studies.

In the Personal Values Questionnaire, the pragmatic mode of valuation is represented by a Successful scale; the ethical–moral mode of valuation is obtained through a Right scale; and the affect or feeling mode of valuation is measured through use of a Pleasant scale.

A combination of primary and secondary modes of valuation was used to operationally define the four conceptual classes of values. Operative values for an individual include those concepts which are rated by him as high in importance and which fit his personal orientation (pragmatic, ethical–moral or affective). Intended values include those concepts rated high in importance but which do not fit his personal orientation, adopted values include those concepts not rated high in importance but which fit his personal orientation. Weak values include those concepts which are neither rated as highly important nor fit the personal orientation of the individual.

Given the aforementioned definitional categories of values, a value score was prepared for each of the 66 concepts on the basis of the responses of each manager by assigning weights as follows: 4 for the operative values, 3 for the intended values, 2 for the adopted values, and 1 for weak values.

Measures of Success

One major problem in dealing with this topic involves the definition and determination of success measurement. What is managerial success and how do we measure it?

Defining and measuring manager success in a study which samples widely across geographical areas, industries, organizations, and type of positions is considerably restricted from that which would be possible if one were studying only managers from one organization.

It was our judgment that the heterogeneous nature of our sample dictated that we define and measure success in terms of objective data that were relatively easy to collect. Success was thus defined as managerial pay relative to age. This success criterion is a relative measure of personal success within the total sample of managers in each country and is not necessarily identical to organizational success or managerial effectiveness. It does, however, provide a cross-organizational measure which is sufficiently accurate for our purposes.

Our procedure for developing a measure of success was as follows: (*a*) Separate managers in each country into eight age categories—20–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55–59, 60 and over. (*b*) Convert the distribution of income received by managers in each age category into a standard score with a mean of zero and a standard deviation of 1.

Thus, a manager in any age category who received a high salary compared to others in his age category would receive a high relative success score (approaching 3.0), while a manager whose salary was low, relative to others in his age group, would receive a low relative success score (approaching −3.0). Each manager, then, had a personal success index between −3.0 and +3.0 depending upon his relative salary position within his age group.

While the success index is useful in relating relative success to value patterns, it does not indicate the extent of success differences in the sample in any meaningful way. Table 1 presents the average salary and the average salary per year of age for the top and bottom 10% of managers in terms of their personal success score.

As shown in Table 1, the managers in the top 10% of our personal success index earn five to six times as much as do the managers in the bottom 10% of our personal success index. Likewise, when one divides salary by age for each individual, the most successful group earns five to six times as much per year of age as does the least successful group. It is clear, then, that our personal success measure represents meaningful variability of con-

TABLE 1

SALARY COMPARISONS OF MOST AND LEAST SUCCESSFUL MANAGERS

Manager	United States		Australia		India		Japan	
	Yearly income (dollars)	Yearly income per year of age	Yearly income (dollars)	Yearly income per year of age	Yearly income (rupees)	Yearly income per year of age	Yearly income (dollars)	Yearly income per year of age
Most successful	41,739	902	25,838	570	81,336	1,944	43,103	848
Least successful	8,045	176	5,500	117	14,616	360	7,500	138
Ratio of high to low	5.2	5.1	4.7	4.9	5.6	5.5	5.8	6.2

siderable magnitude. We are dealing with differences that really make a difference.

Data Analysis

To determine the relationship between personal success and personal value systems for managers, each of the 66 concepts in the Personal Values Questionnaire was examined to determine its degree of relationship with the success index by using the method of double cross-validation (Katzell, 1951). The procedure may be illustrated as follows:

1. Randomly separate the managerial sample into two equal size samples for each of the four countries.
2. Within each subsample, correlate the success score for all individuals with their value scores on each of the 66 concepts.
3. The final scoring keys for each country are developed by selecting those concepts which have the same direction of correlation in both samples and show at least a minimum correlation of .08 in one sample and at least a minimum correlation of .02 in the other sample. Using these two criteria, the product of the two probability values for the occurrence of each concept by chance alone will be less than .05.

TABLE 2

SIGNIFICANT CORRELATIONS FOR CONCEPTS MEASURED BY PERSONAL VALUES QUESTIONNAIRE WITH MANAGERIAL SUCCESS INDEX

Concepts	Countries			
	United States	Australia	India	Japan
High productivity	0	++	+	++
Industry leadership	0	−−	0	+
Organizational stability	−−	−−	0	+
Profit maximization	++	++	++	0
Social welfare	0	−−	−−	−−
Organizational growth	0	++	0	0
Customers	0	++	++	0
Craftsmen	−	++	0	0
Managers	+	++	+	+
Owners	0	++	0	+
My subordinates	++	+	+	+
Labors	0	+	−−	0
My company	+	++	−	0
Stockholders	0	++	0	−
Technical employees	0	0	0	++
Me	+	++	+	−
Labor unions	++	++	+	+
Ability	++	++	++	+
Obedience	−−	−−	−−	0
Trust	−−	−−	−−	0
Agressiveness	+	++	++	++
Loyalty	−−	−−	−−	++
Prejudice	+	+	+	++
Compassion	0	+	−	−−
Cooperation	−−	−	0	0
Tolerance	0	−	−−	0
Conformity	−−	−−	−−	0
Honor	0	−−	0	0
Leisure	−−	−−	−	−
Dignity	−−	−	0	−
Achievement	++	+	++	0
Autonomy	−	++	+	−−
Individuality	0	0	++	0
Influence	++	−−	0	+
Security	−−	−−	−−	−
Power	++	0	+	0
Creativity	+	+	++	0
Success	++	+	+	0
Authority	0	−−	+	+
Caution	−−	0	0	0
Change	++	+	+	0
Competition	0	+	++	++
Compromise	0	−−	−−	0
Conflict	0	0	++	++
Equality	−−	−	−−	−−
Property	0	0	+	−−
Religion	−−	−−	−−	0
Risk	++	−	0	++

Note. ++ indicates a significant positive correlation; −− indicates a significant negative correlation; + indicates a positive correlation; − indicates a negative correlation; 0 indicates no correlation.

The previously mentioned procedure steps were used to develop a special value key for each country, which was not a function of chance differences but was stably related to personal success. The significant correlation results between the concepts measured by the Personal Values Questionnaire and the managerial success index are reported in Table 2. It indicates that of the 66 possible pairs of correlations in each of the four countries, 28 in the Australian sample, 21 in the Indian sample, 12 in the Japanese sample, and 22 in the United States sample were found to be statistically significant. This analysis also shows that there are nine concepts which have the same direction of correlation and have at least one significant relationship across the four countries. Since the probability of this occurring by chance is below .001, it seemed appropriate to develop a common key based on these nine concepts. After the four final scoring keys for each country and the common key for the four countries had been developed, the correlations between the composite value scores of each key and the success index were obtained to indicate the accuracy of the composite value scores of each key for predicting a manager's status on the success index.

RESULTS AND DISCUSSION

Table 3 presents the overall correlations between the success score and the composite score of each value key for the managers in different countries and the total sample. The correlation coefficients in Table 3 were obtained my correlating the success index with their five value scores (as derived from the American, Australian, Indian, Japanese, and common value keys, respectively) for the American, Australian, Indian, Japanese, and total sample. There are three primary results to be noted. First, tailor-made value keys developed for each country manifest the best results. As shown in Table 3, the relationships between personal values and personal success are .32 for American managers, .47 for Australian managers, .35 for Indian managers,

TABLE 3

CORRELATION BETWEEN VALUE SCORES AND THE SUCCESS INDEX FOR MANAGERS IN DIFFERENT COUNTRIES

Concept	No. concepts[a]	Managers				
		American ($n = 878$)	Australian ($n = 301$)	Indian ($n = 500$)	Japanese ($n = 312$)	Total ($n = 1,991$)
Special key based on item analyses of American sample	22	.32*	.30*	.23*	.04	.27*
Special key based on item analyses of Australian sample	28	.18*	*.47**	.26*	.002	.23*
Special key based on item analyses of Indian sample	21	.19*	.29*	*.35**	.03	.23*
Special key based on item analyses of Japanese sample	12	.05	.17*	.19*	*.26**	.13*
Common key among the samples of four countries	9	.23*	.31*	.23*	.15*	.24*

[a] The concepts of each special key are selected by the method of double cross-validation.
* $p < .01$.

and .26 for Japanese managers. These relationships are all statistically significant and are of similar magnitude to the validity coefficients generally reported for predicting manager success by other types of predictors. We view these relationships as solid evidence that value patterns and success are related. Second, Table 3 indicates that when the special value key developed on one country is applied to another country, the validity coefficient shrinks, in some cases to nonsignificant values. This finding implies the danger of applying a validated instrument from one group to another. Third, as shown in the last row of Table 3, the common key value yields significant, yet moderate relationships across the four countries. This indicates that there may be some common elements underlying the relationship between value patterns and success across the different countries.

Since personal value systems of managers are relatively stable and do not change rapidly (Lusk & Oliver, 1972), and since values are related to success, we have explored the possibility of using values as a selection or promotion device in attempting to select individuals who will turn out to be successful. We developed expectancy tables showing the chances out of 100 that a manager with a given value score would be among the top half of managers in terms of the personal success index. (These tables are not shown because of space limitations.) An American manager (for example) who has a value score in the top quintile would have a 72% chance of being included among the top half of managers in terms of success. The expectancy of inclusion decreases to 63%, 47%, 41%, and 34% for decreasing quintile levels of value scores.

From Table 3 and the expectancy tables, it is evident that value patterns predict success and could be used in making selection and placement decisions. However, controversy due to logical, ethical, and legal issues may arise. Persuasive arguments can be made that organizational vitality and adaptation to changing social and technological conditions may come about in large part because of the value mix in an organization. Selecting managers with similar value profiles might result in a static organization. While the full consequences of an individual organization having managers with similar value profiles is not known, it should be noted that the median number of concepts in the four value keys is 21.5 which represents only one third of the total 66 concepts. Use of these value keys in selection and placement would still allow managerial values to differ in the remaining two thirds of the concepts.

A second set of questions about the use of the Personal Values Questionnaire in selection and placement decisions are ethical and legal in nature and center on "fair use" of such an instrument. Table 3 shows that when the special value keys developed from American, Australian, and Indian samples are applied to the Japanese sample and when the Japanese value key is applied to the American sample, the validity coefficients shrink to nonsignificant values. These results are not surprising, however, since subjects from different cultural backgrounds are expected to have different value systems which in turn will result in differential perceptions of managerial success and performance. Although the results indicate that personal value systems do relate to managerial success, they do not show equal validity for managers from different cultures. Thus, without knowing whether the key developed from the American sample is equally valid for other American minority groups, we are uncertain about the fairness of application of the Personal Values Questionnaire to minority groups of American managers.

Other researchers (Brunson, 1970; Kashefi, 1970; Sikula, 1971) have shown that effective and ineffective employees within various organizations have differing values and value systems. Our study also indicates that managerial values have important and similar influences on success under different cultural conditions. Managers from the four different countries are rather similar in terms of the relationship between personal values and success. While the nature of these similarities will be considered in the following section, the reader is asked to note in Table 3 that a common value key which is valid across the four different countries can be developed. It seems likely, therefore, that one could develop a special value key which is valid for different subgroups within a particular population. Since the tailor-made value keys developed for specific countries manifest the best results, we believe different value keys and different expectancy tables should be developed for different subgroups and/or different organizations to promote fairness in employment decisions. Value data also must be integrated with other sources and types of data in order to obtain the optimal utilization of human resources.

Nature of Value Differences between More and Less Successful Managers

Examination of the content of the concepts which are related to success provides additional meaning to the general questions about relationships between values and success.

Looking at the four countries comparatively, one notes that managers from the four countries are rather similar in terms of the relationship between personal values and success. The general pattern in Table 2 indicates that more successful managers have values seated in high productivity, profit maximization, managers, my subordinates, labor unions, ability, aggressiveness, prejudice, achievement, creativity, success, change, competition, and liberalism. Less successful managers have values which emphasize social welfare, obedience, trust, conformity, leisure, dignity, security, conservatism, equality, and religion. Successful managers tend to emphasize pragmatic, dynamic, achievement-oriented values, while less successful managers prefer more static and passive values, the latter forming a framework descriptive of organizational stasis rather than organizational and environmental flux. More successful managers favor an achievement orientation and prefer an active role in interaction with other individuals useful in achieving the managers' organizational goals. They value a dynamic environment and are willing to take risks to achieve organizationally valued goals. Relatively less successful managers have values associated with a static, protected environment in which they take relatively passive roles and often enjoy extended seniority and security in their organizational positions.

In terms of the significant differences that exist among the four countries across the 66 concepts, the most obvious one is that the relationships between the success indexes and the concept "loyalty" are significantly positive for the Japanese managers but are significantly negative for the managers from the other three countries. Another difference shown in Table 2 is that the concept "me" is negatively related to the success index for the Japanese managers but is positively associated with success indexes for the American, Australian, and Indian managers. The behavioral patterns and customs of the Japanese people have been deeply influenced by Confucianism which stresses a rigid relationship within a hierarchically arranged collective society. Members of each collectivity are expected to maintain absolute loyalty and obedience to authority in the fulfillment of their obligations. In view of this cultural background, it is not surprising that more successful Japanese managers place relatively greater emphasis upon "loyalty" and relatively less emphasis upon "me" than do managers in the other three countries.

Although significant correlations between values and success have been found, it seems important to ask whether these correlations represent a superficial and faulty association or an intrinsic relationship. A systematic comparison between the results of our study and other findings in the literature provides evidences of the intrinsic validity of Personal Values Questionnaire for predicting managerial success. A cluster of concepts including high productivity, profit maximization, ability, achievement, and success might first be considered. The positive correlations between these concepts and success imply that an achievement and success motive in the direction of corporate goals is associated with increasing managerial success. Patton (1965) citing his own research and surveys by others noted the relationship between pay and profitability. It might, therefore, be expected that increasing success (as measured in the present study) would be associated with values emphasizing high productivity and a profit maximization. Bray, Grant, and Katkovsky (1967) similarly cite the substantial correlation of career orientation and achievement motivation with managerial salary progress (salary corrected by initial starting salary). McClelland (1953, 1961) also points out that entrepreneurs are primarily driven by the achievement motive. They are interested in productivity, profits, and success because it serves as a measure of their competence. McClelland suggests that achievment motivation precedes individual success and economic growth. Jurgensen (1966) includes productive, intelligent, and enterprising in his list of characteristics most descriptive of successful key executives. The importance of ability is implicit in Williams' (1956) statement that critical executive behaviors include utilization of knowledge and services of subordinates, understanding the capabilities and limitations of his associates, and a major behavioral category concerned with technical competence. Executives are also willing, according to Williams, to work long hours to achieve their assigned objectives (work or achievement primacy rather than a leisure orientation).

A second cluster of values from the present study that are consistent with other findings include change, aggressiveness, creativity, competition, liberalism, social welfare, obedience, trust, conformity, leisure, dignity, equality, conservatism, and religion. More successful managers' values suggest that change, aggressiveness, creativity, competition, and liberalism are behaviorally relevant, while social welfare, obedience, trust, conformity, leisure, dignity, equality, conservatism, and religion are less behaviorally relevant. The implication of more successful managers favoring a dynamic, relatively competitive and initiative environment versus less successful managers' orientation toward a more secure, stable, and convergent environment finds considerable support in the literature. For example, Campbell et al. (1970) include competitive, creative, dedicated, and energetic in their list of personal qualities said to be necessary for managerial effectiveness. Jurgensen (1966) also cites aggressive, self-starting, determined, energetic, and creative as most descriptive of successful executives and cites amiable, conforming, reserved, agreeable, conservative, kindly, mannerly, neat, cheerful, formal, courteous, and modest as least descriptive of successful executives. The previously cited association of achievement motivation with managerial success also is relevant, as are the considerable number of studies noting the direct relationship of dominance and the inverse relation of dependency to managerial success (Bray & Grant, 1966; Bray et al., 1967; Hinrichs, 1967; Wollowich & MacNamara, 1969). Other sources also support the relationships found between the cluster of variables previously noted and managerial success (Dicken & Black, 1965; Laurent, 1962, 1970; Mahoney, Jerdee, & Nash, 1961; Mitchell & Porter, 1967; Porter & Ghiselli, 1957).

A final cluster of concepts encompasses several related to "groups of people" including managers, my subordinates, and labor unions. Managers' perceived behavioral relevance of these concepts increases with their actual success. These concepts are similar to some of those investigated by Rowland (1960), and the results obtained are consistent with the achievement orientation of successful managers noted previously, the group-oriented characteristic of successful managers cited by Campbell et al. (1970), Williams' (1956) dimension of "Relations with Associates," Mahoney's (1966) "Performance–Support–Utilization," and Dicken and Black's (1965) "Ability to Cooperate."

We think the systematic comparisons previously described suffice to show that the correlation between values and success is not a superficial association of nonessential characteristics. There is a real and intrinsic relationship between the level of success achieved by managers and their personal values. The nature of our study (concurrent validation strategy) does not allow any clear interpretation of the causal linkage between values and success. We do not know whether values influence success or whether success influences or changes values among managers. However, f we accept the Rokeach (1968) assumption hat values have strong motivational components as well as cognitive, affective, and behavioral components, it would seem that managers are differentially successful because of their values. The fact that we derived our uccess measure for each age group and then ggregated for all age categories argues gainst there being any strong cumulative ffect of long-term successful experience in he relationship between success and values that we found. Our success measure is not correlated with age and total number of years that a manager has worked. The value keys we developed are minimally related to age in the four countries (*r*s ranging from −.05 to −.17).

SUMMARY

This study examines the nature of the relationship between the personal values of managers and their success as managers for samples of American, Australian, Indian, and Japanese managers. While additional and extended study and analysis of managers in other countries (especially the communist countries) undoubtedly will clarify the functional relationship between values and success, our research evidence suggests the following general findings and conclusions:

1. There is a reasonably strong relationship between the level of success achieved by managers and their personal values. The magnitude of the relationships are indicated by the significant correlation coefficients of .32 for American managers, .47 for Australian managers, .35 for Indian managers, and .26 for Japanese managers. We view these correlations as real and intrinsic relationships rather than superficial and faulty associations.
2. It is evident that value patterns are predictive of managerial success and could be used in selection and placement decisions. A number of legal and ethical issues must be considered, however, before such use is justified.
3. Although there are country differences in the relationships between values and success, the findings across the four countries are quite similar.
4. The general pattern emerging from the study indicates that more successful managers appear to favor pragmatic, dynamic, achievement-oriented values, while less successful managers prefer more static and passive values, the latter forming a framework descriptive of organization stasis rather than organizational and environmental flux. More successful managers favor an achievement orientation and prefer an active role in interaction with other individuals instrumental to achievement of the managers organizational

goals. Less successful managers have values associated with a static and protected environment in which they take relatively passive roles.

REFERENCES

Allport, G. W., Vernon, P. E., & Lindzey, G. *Study of values: A scale for measuring the dominant interests in personality.* (3rd ed.) New York: Houghton Mifflin, 1960.

Bray, D. W., & Grant, D. L. The assessment center in the measurement of potential for business management. *Psychological Monographs,* 1966, **80**(17, Whole No. 625).

Bray, D. W., Grant, D. L., & Katkovsky, W. Contributions of projective techniques to the assessment of management potential. *Journal of Applied Psychology,* 1967, **51**, 226–232.

Brunson, R. W. A behavioral case study of a miniconglomerate. Unpublished doctoral dissertation, Michigan State University, 1970.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., and Weick, K. E. *Managerial behavior, performance, and effectiveness.* New York: McGraw-Hill, 1970.

Dicken, C. F., & Black, J. D. Predictive validity of psychometrists evaluations of supervisors. *Journal of Applied Psychology,* 1965, **49**, 34–47.

England, G. W. Personal values systems of American managers. *Academy of Management Journal,* 1967, **10**, 53–68.

England, G. W. Personal values systems of Australian managers. Mimeographed report, Industrial Relations Center, University of Minnesota, 1972.

England, G. W., Agarwal, N. C., & Dhingra, O. P. *Personal value systems of Indian managers.* Minneapolis: University of Minnesota Press, 1973.

England, G. W., & Koike, R. Personal value systems of Japanese managers. *Journal of Cross-Cultural Psychology,* 1970, **1**, 21–40.

Ghiselli, E. E. *The validity of occupational aptitude tests.* New York: Wiley, 1966.

Hinrichs, J. R. Comparison of "real life" assessments of management potential with situational exercises, paper and pencil ability tests, and personal inventories. *Journal of Applied Psychology,* 1967, **53**, 425–432.

Jurgensen, C. D. Report to participants on adjective wordsort. Unpublished report, Minneapolis Gas Company, 1966.

Kashefi-Zihajh, M. An empirical investigation of the relationship between value systems and organizational effectiveness. Unpublished doctoral dissertation, Michigan State University, 1970.

Katzell, R. A. Cross-validation of item analyses. *Educational and Psychological Measurement,* 1951, **11**, 16–22.

Laurent, H. The validation of aids for the identification of managerial potential. Standard Oil Company of New Jersey, 1962.

Laurent, H. Cross-cultural cross-validation of empirically validated tests. *Journal of Applied Psychology,* 1970, **54**, 417–423.

Lusk, E. L., & Oliver, B. L. *The impact of organizational interactions on the American manager's personal value system.* The Wharton School, University of Pennsylvania, 1972.

Mahoney, T. A. Criteria of organizational effectiveness. Mimeographed report, Industrial Relations Center, University of Minnesota, 1966.

Mahoney, T. A., Jerdee, T. H., & Nash, A. N. *The identification of management potential.* Dubuque, Iowa: W. C. Brown, 1961.

McClelland, C. D., Atkinson, J. W., Clark, R. A., & Lowell, E. L. *The achievement motive.* New York: Appleton-Century-Crofts, 1953.

McClelland, D. C. *The achieving society.* Princeton, N.J.: Van Nostrand, 1961.

Mitchell, V. F., & Porter, L. W. Comparative managerial role perceptions in military and business hierarchies. *Journal of Applied Psychology,* 1967, **51**, 449–452.

Osgood, E. C., Suci, G. J., & Tannenbaum, P. H. *The measurement of meaning.* Urbana: University of Illinois Press, 1957.

Patton, A. Deterioration in top executive pay. *Harvard Business Review,* 1965, **32**, 106–118.

Porter, L. W., & Ghiselli, E. E. The self-perceptions of top and middle management personnel. *Personnel Psychology,* 1957, **10**, 397–406.

Rokeach, M. *Beliefs, attitudes and values.* San Francisco: Jossey-Bass, 1968.

Rowland, V. K. *Managerial performance standards.* New York: American Management Association, 1960.

Sikula, A. F. Values and value systems: Importance and relationship to managerial and organizational behavior. *Journal of Psychology,* 1971, **78**, 277–286.

Wollowich, H. B., & McNamara, W. J. Relationship of the components of an assessment center to management success. *Journal of Applied Psychology,* 1969, **53**, 348–352.

Williams, R. E. A description of some executive abilities by means of the critical incident techniques. Unpublished doctoral dissertation, Columbia University, 1956.

(Received June 11, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 4, 420-423

INFLUENCE OF DOGMATISM, RISK-TAKING PROPENSITY, AND INTELLIGENCE ON DECISION-MAKING STRATEGIES FOR A SAMPLE OF INDUSTRIAL MANAGERS

RONALD N. TAYLOR[1]
University of British Columbia

MARVIN D. DUNNETTE
University of Minnesota

The relative influence of dogmatism, risk-taking propensity, and intelligence on strategies adopted in a decision-making exercise was investigated for a sample of industrial managers. Findings of prior studies, that dogmatism was associated with a decision-making strategy characterized by rapid and confidently held decisions following limited information search, were supported. High propensity for risk was associated with an information-seeking strategy involving rapid decisions made on the basis of little information, but involved deliberate information processing. Intelligence was positively correlated with efficiency in processing information, accurate choices, and cautiousness in changing decisions in the face of adverse consequences.

The influence of psychological attributes of decision makers on the strategies they adopt in decision-making tasks has received considerable attention in the research literature (e.g., Bourne, Ekstrand, & Dominowski, 1971; Lanzetta & Kanareff, 1962; MacCrimmon & Taylor, in press). Dogmatism has been singled out for intensive investigation, perhaps because of its central role in theories of decision making (e.g., Rokeach, 1960), and some tentative conclusions have been drawn to explain its impact on decision-making strategies. Dogmatism generally has been found to be associated with an approach to decision making characterized by rapid decisions following limited information search and yet confidently holding to the resulting decisions (e.g., Long & Ziller, 1965; Pruitt, 1961).

A decision maker's risk-taking propensity and his intelligence probably also figure prominantly in his decision-making strategies, but the influence of these psychological attributes on such strategies has not been systematically investigated. Under conditions of relatively costless information seeking and increased probability of attaining correct decisions as additional relevant information is acquired, high-risk-taking propensity would be expected to be associated with decision-making strategies stressing limited information seeking and rapid decision times. The influence of the intelligence of decision makers on his decision-making approaches has generated surprisingly little research; it logically should influence the evaluative aspects of decision making. The purpose of the present study was to investigate the impact of these three psychological attributes on the information-seeking and decision-making strategies used by a sample of industrial managers for making a job-related decision.

METHOD

Measures of dogmatism, risk-taking propensity, and intelligence were administered to a sample of 79 male, line managers; ranging from foremen to factory managers and representing eight units of a heavy manufacturing company. The subjects then participated in an individually administered decision-making exercise. The research was conducted while the subjects were participating in a management development program at the corporate headquarters, and data were obtained for all participants in three sessions of the program.

Instruments

Psychological attributes investigated were dogmatism, measured by the Dogmatism scale (Rokeach, 1960); risk-taking propensity, measured by the Choice Dilemmas Procedure (Kogan & Wallach, 1964); and intelligence, measured by the Wesman Personnel Classification Test (Wesman, 1965). The

---

[1] Requests for reprints should be sent to Ronald N. Taylor, Faculty of Commerce, University of British Columbia, Vancouver 8, British Columbia, Canada.

The authors are indebted to Larry Cummings and George Huber for helpful comments on an earlier draft of this article.

decision-making behaviors of amount of information sought, time required to reach a decision, information-processing rate, decision accuracy, decision confidence, and decision flexibility were measured by the Personnel Decision Simulation developed by Taylor (1970).

Personnel Decision Simulation. Information-seeking and decision-making strategies were measured by the Personnel Decision Simulation. The Personnel Decision Simulation presents the subject with a simulated managerial decision requiring him to play the role of a business manager while making a promotion decision. The subject selects, on the basis of information that an actual managér would have available to him, one of three hypothetical salesmen for promotion to sales manager. The Personnel Decision Simulation departs from experimental tasks typically used to investigate decision-making performance by emphasizing systematic observation and objective measurement of processes involved at the various stages of decision making. The decision processes measured by the Personnel Decision Simulation are (*a*) information amount, the number of information items sought; (*b*) decision time, the time required to reach a decision; (*c*) processing rate, the rate at which information units are processed; (*d*) decision accuracy, correctness of decision choices; (*e*) decision confidence, self-rated confidence in the correctness of decisions; and (*f*) decision flexibility, the number of adverse consequences of a decision which would cause the decision to be changed (high score indicates flexibility). Scores on the Personnel Decision Simulation are based on two decision problems. The two problems were equivalent forms of the Personnel Decision Simulation administered with a one-week interval between forms and total scores were the mean scores for the two forms. For any subscore (e.g., "time") the correlation between scores for the two decision problems was calculated and entered in the Spearman-Brown formula. Means, standard deviations, and reliabilities for the variables used in this study are given in Table 1.

TABLE 1

MEANS AND STANDARD DEVIATIONS FOR PERSONNEL DECISION SIMULATION SCORES AND PSYCHOLOGICAL ATTRIBUTES FOR 79 MALE LINE MANAGERS

Variable	$\bar{X}$	*SD*	Reliabilities[a]
Personnel Decision Simulation scores			
Amount of information	138.04	57.26	.98
Processing rate	3.70	1.28	.96
Decision time	38.49	14.11	.96
Decision accuracy	.72	.39	.68
Decision confidence	3.08	.81	.78
Decision flexibility	7.56	2.74	.93
Psychological attributes			
Intelligence	29.79	10.08	
Risk-Taking propensity	40.05	7.67	
Dogmatism	19.08	27.65	

[a] Coefficients for two forms of the Personnel Decision Simulation corrected by Spearman-Brown formula.

TABLE 2

CORRELATIONS BETWEEN PERSONNEL DECISION SIMULATION SCORES AND PSYCHOLOGICAL ATTRIBUTES

Personnel decision simulation scores	Dogmatism	Risk-taking propensity	Intelligence
Amount of information	−.16	−.37**	−.16
Processing rate	.18	−.21*	.39**
Decision time	−.34**	−.29**	−.10
Decision accuracy	.23*	−.04	.30**
Decision confidence	.27*	.11	.09
Decision flexibility	−.09	−.15	−.45**

* $p < .05$.
** $p < .01$.

Analysis

The bivariate correlation coefficients between dogmatism, risk-taking propensity, intelligence, and the six scores from the Personnel Decision Simulation were examined to determine both the nature of the psychological traits and the relationships between these traits and performance in the decision-making exercise.

RESULTS AND DISCUSSION

Intercorrelations Among Decision-Maker Attributes

The intercorrelations were insignificant between risk-taking propensity and both intelligence (.09) and dogmatism (.08). The only significant relationship was between intelligence and dogmatism (.29), with dogmatic subjects tending to be more intelligent. These relationships are so low it appears that only a small part of the influence of dogmatism on decision-making performance found by prior studies could have been attributed to either the intelligence or the risk-taking propensity of dogmatic decision makers.

Correlations Between Psychological Attributes and Personnel Decision Simulation Scores

The correlation coefficients between psychological attributes and Personnel Decision Simulation scores are shown in Table 2.

Dogmatism. Correlations between the Rokeach Dogmatism scale and the Personnel Decision Simulation scores that were in line with findings from previous studies were the significant negative correlation between dogmatism and decision time and the positive correlation between dogmatism and decision confidence (e.g., Long & Ziller, 1965). As could be expected from the correlations between the variables and the data in Table 2, the significant correlation between the decision accuracy and dogmatism becomes insignificant (.14) when intelligence is partialed out.

Risk-Taking Propensity. Risk-taking propensity, as measured by the Choice Dilemmas Procedure, was found to be negatively associated with both the time required to reach a decision and the amount of information upon which the decision was based. For the Personnel Decision Simulation this is a risky strategy, because the probability of a correct decision increases as additional relevant information is acquired. There was, however, a significant tendency for risk-taking propensity to have a negative association with the *rate* at which the Personnel Decision Simulation information was processed. Apparently, high-risk takers made more rapid decisions based on less information, but tend to process each item of information more *slowly*. This finding is at variance with a conceptualization of risk taking as disregard for the information base in a willingness to be exposed to an uncertain decision environment. It suggests that, although risk-prone decision makers reach rapid decisions by the expedient of restricting their information search, they give careful attention to the information they acquire.

Understanding of the nature of risk-taking propensity is far from complete. Attempts to determine the meaning and dimensionality of risk taking as a construct (e.g., Wallach & Kogan, 1961; Slovic, 1962) have found the correlations among instruments proported to measure risk taking to be generally low. Such results led Slovic (1964) to conclude that either very few of these measures tap general risk-taking propensity or that such propensities do not exist. Kogan and Wallach (1967) suggested that until careful empirical work is carried out in these areas, attempts to conceptualize various judgmental and cognitive processes in decision and risk terms must be considered premature. One purpose of the present study was to further clarify the interpretation of the Choice Dilemmas Procedure as a measure of risk-taking propensity, both in the light of its relationships with dogmatism and intelligence and its influence on decision-making performance.

Intelligence. Intelligence showed significant positive correlation with a tendency to process information quickly and to make accurate decisions. It was also found to be negatively associated with decision flexibility. Apparently the major contribution of intelligence to decision-making performance is in the efficiency with which information was processed and accuracy in choosing a high-quality decision alternative. Intelligence appears to have little influence on decision confidence, but it is important in evaluating consequences of a decision to determine which ones are sufficiently serious to warrant changing the decision. In each of these decision processes, the influence of intelligence seemed to involve evaluation of information, integrating it for making a choice and judging the import of adverse consequences. Intelligence apparently had little influence on the more idiosyncratic processes of predecisional information search strategies and postdecisional confidence. Further studies using measures of general mental ability and extending the investigation to include specific mental abilities, are needed to understand more fully the role of intelligence in decision making.

Investigation of the psychological correlates of information-processing and decision-making performance appears to be a fertile area for future research into the cognitive and motivational processes involved in managerial decision making. Further investigations are needed to specify more completely the influence of decision-maker characteristics on performance in decision problems. This study indicates that the Personnel Decision Simulation is a potentially useful tool for achieving those ends.

REFERENCES

Bourne, L. E., Jr., Ekstrand, B. R., & Dominowski, R. C. *The psychology of thinking*. Englewood Cliffs, N.J.: Prentice-Hall, 1971.

Kogan, N., & Wallach, M. A. *Risk taking: A study in cognition and personality*. New York: Holt, Rinehart & Winston, 1964.

Kogan, N., & Wallach, M. A. Risk-taking as a function of the situation, the person and the group. *New [illegible]ctions in psychology III.* New York: Holt, [illegible]hart & Winston, 1967.

Lanzetta, J. T[illegible] Kanareff, V. T. Information cost, amount of [illegible]ff, and level of aspiration as determinants [illegible]formation seeking and decision-making. *Be[illegible]ral Science,* 1962, **7**, 459–473.

Long, B. H., [illegible]iller, R. C. Dogmatism and predecisional [illegible]mation search. *Journal of Applied Psychology,* [illegible]5, **49**, 376–378.

MacCrimmon, [illegible] R., & Taylor, R. N. Decision-making and [illegible]blem-solving. In M. D. Dunnette (Ed.), *Ha[illegible]k of industrial and organizational psychology.* [illegible]cago: Rand McNally, in press.

Pruitt, D. G. [illegible]rmational requirements in making decisions. *[illegible]ican Journal of Psychology,* 1961, **74**, 433–4[illegible]

Rokeach, M. [illegible] *open and closed mind.* New York: Basic Book[illegible], [illegible]960.

Slovic, P. Convergent validation of risk-taking measures. *Journal of Abnormal and Social Psychology,* 1962, **65**, 68–71.

Slovic, P. Assessment of risk-taking behavior. *Psychological Bulletin,* 1964, **61**, 220–233.

Taylor, R. N. *The development and evaluation of an instrument to examine individual differences in managerial decision-making styles.* Unpublished doctoral dissertation, University of Minnesota, 1970.

Wallach, M. A., & Kogan, N. Aspects of judgment and decision-making: Interrelationships and changes with age. *Behavioral Science,* 1961, **6**, 23–36.

Wesman, A. G. *Personnel Classification Test manual, 1965 revision.* New York: Psychological Corporation, 1965.

(Received May 29, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 4, 424-432

QUALITY OF DECISIONS:

INDIVIDUALS VERSUS REAL AND SYNTHETIC GROUPS

BERTRAM SCHONER[1] AND GERALD L. ROSE

University of Iowa

G. C. HOYT

Simon Fraser University

The quality of decisions on three economic problems was compared for individuals, two forms of real five-man groups, and three synthetic group decision rules. Both types of real groups outperformed individuals. Real groups with no previous exposure to the problems made better decisions than did groups whose members had previously made individual decisions on the same tasks and made decisions superior to a plurality synthetic decision rule, but inferior to the "best man" or Lorge and Solomon's Model A rule. The relationship between decision quality and riskiness of decision alternatives varied in the three decision tasks. Real groups with no prior exposure to the task responded to the suggested risk norm more appropriately than did other decision-making units.

This article reports the results of an experiment on the effects of individual and alternative modes of face-to-face group decision making on decision quality. Identical economic decisions were made by individuals, by groups who had first considered and reached decisions on the problems individually, and by groups who had no prior exposure to the problems. In addition, the latter groups, subsequent to the group decisions, made individual decisions. Thus this investigation differs from the bulk of previous problem-solving studies principally in that two types of groups were employed, and the problems required economic decisions under uncertainty.

Much of the early work comparing group versus individual performance on problem-solving tasks tended to show the superiority of groups over individuals (Lorge, Fox, Davitz, & Brenner, 1958; Lorge & Solomon, 1955). Subsequent studies, however, tended to yield mixed results in comparing performance, indicating that under some conditions the group setting inhibits performance, while in others performance may be enhanced (Dunnette, Campbell, & Jaastad, 1963; Hall, Mouton, & Blake, 1963; Taylor, Berry, & Block, 1958; Tuckman & Lorge, 1962). In a study by Campbell (1968) the quality of group solutions was clearly inferior to the scores of even the average individuals. As Campbell points out, however, "such a finding is dependent upon the type of problem used and the manner in which it was scored [p. 209]."

Another branch of literature relevant to the present study has shown similar results—that of individual risk taking and, in particular, the so-called "risky shift" phenomenon. Hereto, risk-taking propensity has been shown to be highly situation and task specific (Slovic, 1962), and especially the enhancement of risk-taking propensity under group conditions (the so-called "risky shift" effect) has been shown to be highly specific to the task involved. For a critical review of the "risky shift" literature, see the special supplement of the December 1971 *Journal of Personality and Social Psychology*, in which it is pointed out that the risky shift phenomenon is really specific to *some* of the dilemmas of choice items originally used by Stoner (1961). However, for evidence that the phenomenon operates, at least in part, for simulated investment portfolio decisions, see Deets and Hoyt (1970).

Despite such mixed results, these studies might well have interesting implications for economic decision-making organizations if the phenomena in question could be shown to hold for economic decision problems. Over a decade ago Simon (1960) made a strong appeal for testing the implications of just such behavioral phenomena for the theory of decision making in economics. The present study constitutes one attempt to meet this objective.

---

[1] Requests for reprints should be sent to Bertram Schoner, who is now at Department of Economics and Commerce, Simon Fraser University, Burnaby, British Columbia, Canada.

Unlike most previous studies of individual and group decision making that used noneconomic decision tasks, the present study relied upon statistical decision theory to define decision quality. Each subject and group made three business decisions. In each decision each alternative was associated with clearly specified possible events, a probability distribution on these events, and a payoff associated with each event. The attitude toward risk was imposed in that for each problem the decision maker (or group) was to make the decision in the context of a role in a firm. The description of the firm indicated whether or not he was to be risk averse or risk neutral.

The decision criteria below follow from statistical decision theory:

1. If one alternative has at least as high a payoff as that of a second alternative for every possible event, and if there exists at least one event for which the first alternative has a higher payoff, the first alternative *dominates* and is superior to the second.
2. If the decision maker (or group) should be risk neutral, alternatives possessing higher expected values are superior to alternatives possessing lower expected value. In the context of utility theory, risk neutrality implies a linear utility function over the range of payoffs of the alternatives.
3. If the decision maker or group should be risk averse (a) higher expected value is superior to lower expected value, given equal variance ("variance" is used in the statistical sense as a measure of dispersion of a random variable); (b) lower variance is superior to higher variance, given equal expected value; (c) a higher expected value-lower-variance alternative is superior to a lower expected value-higher-variance alternative.

Criterion 3 merits some discussion. By this criterion, expected value and variance are the only decision parameters. This has been the tradition in much of the normative decision theory since Markowitz (1959). In addition, if a utility function is approximately quadratic over the range of interest, expected utility maximization is achieved while only considering expected value and variance (Krantz, Luce, Suppes, and Tversky, 1971, pp. 398–399).

As Dion, Miller, and Baron (1970) and Barron and Mackensie (1971) have noted, the concept of risk has not been well defined in the psychological literature. Most frequently, in both the Choice Dilemma Questionnaire (Kogan & Wallach, 1964) and expected value (e.g., Bem, Wallach, & Kogan, 1965) approaches, the focus is upon the probability of loss associated with each alternative. Slovic (1962) also found that in gambling situations perceptions of the riskiness of a gamble are most closely associated with the probability of loss. In contrast, the present study emphasizes the expected value-variance analysis of risk. This redefinition is trivial if one is concerned primarily with explaining group processes such as the process of "shifting" attitudes (see Dion et al., 1970). However, it is critical if one is concerned with identifying which factors affect economic decisions. The distinction between the probability of loss and variance definitions of risk is particularly important in binominal outcome situations like Slovic's (1962) in which payoffs for each event are fixed and probabilities vary. In binominal situations, the maximum variance obtains when each outcome is equally probable (i.e., probability of loss = probability of gain = .5) regardless of payoff values. Variance decreases to a mimimum (zero) as the probability of loss approaches either zero or one. Thus a high probability of loss is not as risky (in the variance sense) as an intermediate probability of loss.

Method

Each subject (an undergraduate student in business administration) participated in making six decisions in one of two randomly assigned treatments. Ninety-five subjects first reached three decisions as individuals and were then randomly assigned to one of 19 five-man groups to make an additional three decisions on identical problems. These 95 subjects are referred to as having a I–G treatment. Another

90 subjects considered the same decisions but in the reverse (G–I) order. As before, the 90 G–I subjects were assigned to their five-man decision-making groups in a random manner. (In both group conditions no leaders were appointed. Rather, leadership was permitted to emerge.)

The decision problems are described as follows.

Problem 1

The first decision problem is shown in the present context, as the "Holigraph Corporation" problem, in the form in which it was presented to subjects making individual decisions. In order not to create a bias by identifying plant locations, nonsense syllables were used (Cohen & Hensel, 1956) instead of letters. The order of presentation of plant locations was randomized. For group decisions, the decision-making unit was identified as the "Executive Committee" rather than the "President." In all other respects, the problems were identical.

Holigraph Corporation

> As the President of the Holigraph Corporation, you must decide in which of four regions (coded below to maintain secrecy from competitors) you will build a plant to produce a new communications device. Table 1 shows the resulting rate of return on investment for each combination of plant location and degree of market acceptance.
>
> Holigraph's market research indicates that the probability of very high consumer acceptance is 10 percent, the probability of moderately high acceptance is 30 percent, the probability of very low acceptance is 20 percent, and the probability of low acceptance is 40 percent.
>
> Holigraph has historically been conservative in its investment policy, and you have no reason to believe that the Board of Directors has changed this orientation.
>
> In which location would you build? Briefly state your reasons for the decision you reached on the reverse side of this page.

The application of the three decision rules (described previously) to the expected values and variances of the locations yields a quality ranking of (*a*) ZAJ, (*b*) ZOJ, (*c*) ZEJ, and (*d*) ZUJ.

The "conservative investment policy" of the board should imply, for two investments of equal expected value, that the smaller variance alternatives is superior. (See Criterion 3 previously mentioned.)

Although it is an unsettled question as to which alternative is riskier if one has a higher expected value and the other a lower variance, we can assert that: (*a*) a dominated alternative is riskier; and (*b*) of two alternatives possessing equal expected value, the higher variance alternative is riskier. (See our earlier discussion.) Applying these criteria to the Holigraph problem yields the identical risk ranking as by focusing on probability of lowest outcome; that is, there is no ambiguity in assessing risk. The resulting ranking by risk is (*a*) ZUJ, (*b*) ZEJ, (*c*) ZOJ, and (*d*) ZAJ. It should be noted that the risk ranking is in exactly the opposite order to the quality ranking.

Problem 2

The second task is shown in the present context, as the "Zirolex I" problem, in the form received by subjects.

Zirolex I

> As the Vice-President of Marketing for the General Chemical Products Corporation, a very large producer of industrial chemicals, you must decide whether or not to produce one last batch of Zirolex (a chemical compound used in the production of paint).
>
> Last year General Chemical made $214 million after taxes. Although the profit margin of Zirolex has been declining in recent years, last year sales of Zirolex produced a contribution to profit of $80,000.
>
> The Zirolex producing equipment was originally purchased for $300,000 ten years ago, but is now fully depreciated. The equipment could be sold for scrap at any time, netting the company $10,000 after miscellaneous costs associated with the removal of the equipment.
>
> You know that one of your competitors is

TABLE 1

RATE OF RETURN ON INVESTMENT AND DERIVED MEASURES FOR HOLIGRAPH LOCATIONS

Plant location[a]	Degree of market acceptance				Quality rank[bc]	Expected value[b]	Variance[b]	Risk rank[bc]
	% very high	% moderately high	% low	% very low				
ZEJ	15	11	9	8	3	10	4	2
ZAJ	20	16	14	13	1	15	4	4
ZUJ	29	17	4	2	4	10	75	1
ZOJ	22	16	15	10	2	15	10+	3

[a] Nonsense syllables were used for plant locations.
[b] Not given to subjects.
[c] Note that ZAJ and ZOJ dominate ZEJ.

developing a new compound which will make Zirolex obsolete when it is brought out. Your production and marketing experts studied the problem carefully and have reached a consensus that there is a fifty-fifty chance that this competitor will bring out the compound within two months.

If you decide to produce one more batch of Zirolex, because of scheduling difficulties with the packaging department, the batch would have to be produced next week. You are able to produce one of two batch sizes. If a large batch (4 tons) were to be produced, and if the competitor's new compound were marketed within the next two months, most of the batch would have to be sold off at drastically reduced prices. This would result in a loss of $20,000 on the batch. On the other hand, if the competitor's new product were not brought out within the next two months, you would expect to make a profit of $50,000 on the batch.

If a small batch (2 tons) were to be produced, and if the competitor's new product were to be introduced within the next two months, the expected loss would be $5,000 on sales of the batch. If the competitor's new product were not marketed within two months, you expect a contribution to profits of $25,000 on sales of the batch. Of course, you also have the option of producing no Zirolex at all next week.

What will you do? Why?

An optimal strategy for a large company making numerous small decisions would be to maximize expected value and be risk neutral (i.e., variance is irrelevant) on each decision, since the firm has the law of large numbers working for it. Thus, the optimal alternative is to produce the large batch, at an expected value of $15,000, and the worst alternative is to produce nothing.

The previously outlined criteria for ranking alternatives with respect to riskiness are inadequate for Zirolex I. Consider the following extension of the risk criteria. *For nondominated alternatives, that alternative possessing the highest variance is riskiest.* Accepting this definition again insures that the variance and probability of loss notions of risk yield identical rankings of alternatives in Zirolex I problem. The large batch is riskiest followed in order by small batch and no batch.

Problem 3

A third task is shown in the following display, as the "Zirolex II" problem, in the form received by subjects.

Zirolex II

Just this morning, you have received a telephone call from John Ransome, the President of a reputable research firm. Ransome told you that your competitor has recently marketed a small quantity of the new compound through a wholesaler in order to have the compound tested. Ransome's research firm has obtained a pound of the substance. Ransome offered to test the compound, and tell you *positively* within 24 hours, whether or not product development has proceeded to the point where the compound will be marketed within two months. You know that your competitor is waiting only for technical development of the compound to be completed before marketing it. Therefore, you are sure that the information you would receive from Ransome's research firm would allow you to know with certainty whether or not the new compound will be marketed within two months.

Ignore any tax and ethical considerations in answering the following question: What is the most you would be willing to pay Ransome's research firm for the information on your competitor's compound? Why?

The expected value of the optimal act (large batch) under uncertainty is $15,000. If perfect information indicated that the competition will market his product within two months, the optimal batch size is zero for zero payoff. Conversely, if it were known with certainty that the competitor's new product will not be marketed within two months, the optimal batch size is large, for a payoff of $50,000. The analysis indicates a .5 probability that the competitor's product will be marketed within two months.

Thus the expected value under certainty (i.e., if the information is purchased) is $.5 \times \$50,000 + .5 \times \$0 = \$25,000$.

This is $10,000 increase over the expected value which can be obtained without information. Hence, a rational decision maker should be willing to pay up to $10,000 for the information.

Procedure

At the beginning of the experimental session, subjects were instructed that they would be asked to make a series of decisions similar to those encountered in actual business situations. They were asked to act as they would if they actually occupied the role assigned to them (e.g., President in the individual condition and member of the Executive Committee in the group condition). They were given no training in statistical decision theory during the experiment nor were they instructed to do their best in any sense except in attempting to behave as they actually would in the roles they were asked to assume.

Half of the subjects in each treatment received Problem 1 (Holigraph) first; the other half received Problems 2 and 3 (Zirolex I and Zirolex II) first. (Problems 2 and 3 were necessarily paired by the nature of the issues to be decided.) The order of the three problems was preserved in the individual and group conditions for each subject. Within each task the order of presentation of alternatives was randomly assigned for each decision-making unit. No significant order effects were observed.

Finally, at the end of the experiment, the subjects were debriefed, paid, and excused.

No time limits were placed upon subjects in either the individual or group conditions. When all subjects within a given (e.g., individual) condition were finished they immediately began their consideration of the additional three decisions in the new (e.g., group) condition without receiving any evaluation of their first three decisions. After recording each decision, subjects were asked to write a very brief justification for the decision and rated their confidence in the decision on a 7-point Likert-type scale. After finishing the second set of three decisions, subjects were given a postexperimental questionnaire designed to obtain sociometric measures. (Since no significant results were obtained upon analysis of these latter data, the results are not described in this article.)

RESULTS

Quality of Decisions

The relative quality of decision-making performance for each condition is analyzed for each task in this section. The quality of decisions in each condition is then compared to the predicted decisions obtained by three pooling rules used to create synthetic group decisions from individual decisions: (*a*) plurality, (*b*) best man's decision, and (*c*) Lorge and Solomon's (1955) Model A.

As Table 2 shows, the groups in the G–I treatment clearly showed superior performance on the Holigraph problem, selecting the optimal location two thirds of the time, compared to less than one third for the I–G groups, and only 16.8% for the individuals in the I–G treatment. (If one chooses to consider only the expected value criterion, thus making locations ZOJ and ZAJ equally desirable, the same relative performance obtains.) Although the academic experience of the subjects was such that it would be expected, ideally, that all subjects should possess sufficient sophistication to calculate expected value, not all did so (or did not perceive the relevance of the technique). To minimize risk, on the other hand, it was not necessary to calculate the variance of each payoff distribution. A simple perception of the range of payoffs is a sufficient estimate of the variance in this matrix.

TABLE 2

QUALITY OF DECISIONS BY GROUPS AND INDIVIDUALS ON PROBLEM 1

Quality rank	Location[a]	Individual then group treatment (I–G)		Group then individual treatment (G–I)	
		% individuals ($n = 95$)	% groups ($n = 19$)	% groups ($n = 18$)	% individuals ($n = 90$)
1	ZAJ	16.8	31.6	66.7	61.1
2	ZOJ	42.1	57.9	33.3	35.6
3	ZEJ	4.2	0	0	1.1
4	ZUJ	36.8	10.5	0	2.2

Note. $X^2 = 64.5$, $p < .001$ (omitting ZEJ because of low-expected-cell frequency).

[a] Nonsense syllables were used for plant locations.

Individuals in the G–I treatment tended to give the same responses as their group had given. Only three of the 90 individual subjects in the G–I treatment abandoned the location chosen by the group.

Table 3 shows that in the Zirolex I task the majority of groups and individuals, under all treatments, tended to prefer the medium-risk, medium-expected-payoff alternative, that is, to produce a small batch. Nevertheless, inspection of Table 3 shows the pattern of G–I group decisions to be superior to that of either I–G groups or I–G individuals, with I–G groups marginally superior to I–G individuals. The G–I individuals were marginally inferior to G–I groups, with very few shifts from the group decision. Thus, the relative quality performances of the various decision modes observed in Problem 1 is essentially replicated in Problem 2.

Contrary to the results of some previous studies (e.g., Green, Halbert, & Fitzroy, 1967), the subjects in this study rather consistently undervalued the information which they could purchase in the Zirolex II task.

In this instance, individuals in the G–I treatment achieved the highest proportion of the optimal $10,000 decision (the few switches tended to be in the optimal direction), followed by G–I groups, then by I–G groups, with the I–G individuals again making the poorest showing. These differences, while consistent with the results of Problems 1 and 2, do not achieve statistical significance.

In order to gain insight as to whether the relative superiority of real groups over individuals was due to statistical pooling or to the dynamics of face-to-face interaction, synthetic group decisions were derived. The composition of synthetic decision groups was

TABLE 3

QUALITY OF DECISIONS BY GROUPS AND INDIVIDUALS ON PROBLEM 2

Batch size	Expected value	Risk (variance)	Individual then groups treatment (I–G)		Group then individual treatment (G–I)	
			% individuals ($n = 95$)	% groups ($n = 19$)	% groups ($n = 18$)	% individuals ($n = 90$)
None	$0	0	10.5	0	0	5.6
Small	$10,000	Medium	73.7	84.2	66.7	64.4
Large	$15,000	High	15.8	15.8	33.3	30.0

Note. Kruskal-Wallace is significant at .05 level. $H = 8.3$, $p < .05$.

identical to that of the I–G groups so that comparisons of the I–G groups' natural and synthetic decisions involve the same subjects.

Three pooling rules were employed to generate the synthetic decisions. First, a plurality vote (with ties resolved in favor of the best alternative) rule was used to combine the individual decisions of I–G subjects. The plurality rule (two or more individual decisions in the group being optimal, no majority for a nonoptimal decision, and with ties resolved in favor of the optimal decision) represents a stronger test than a majority rule. (In fact, in no instance would the majority rule have resulted in an optimal decision.) A second set of synthetic group decisions was then obtained by assuming that the group's synthetic decision was the same as the best individual decision. Finally, the Lorge and Solomon (1955) Model A was used to predict group decisions on the basis of the individual data.

If real groups outperform the plurality vote synthetic decision rule, then one might infer that some positive contribution can be ascribed to face-to-face relationships. Table 4 shows that both I–G and G–I groups tended to reach more optimal decisions than did plurality vote synthetic groups. However, except for G–I groups in Problem 1, neither type of real group reached as many optimal decisions as would be expected by the best-man synthetic criterion. The latter results are typical of studies surveyed by Davis (1969). Apparently face-to-face interaction is not a sufficient condition to assure adoption of the best individual's decision, but it is more than a mere voting process which selects among several individual viewpoints.

The same conclusion is suggested by application of the Lorge and Solomon (1955) Model A to the present data. Their model asserts that the probability that a group will make an optimal decision is $1 - (1 - P)^r$ where r is the number of members of the group and P is the likelihood that an individual will make an optimal decision. This criterion is similar to the best-man-out rule in that it is assumed that if an individual reaches an optimal decision, this decision would be adopted by the (artificial) group of which he is a member. One may use the frequency of optimal decisions in the individual I–G results as a measure of P. Consistent with previous studies (Davis, 1969, p. 46) Model A overpredicts group performance in this study with one notable exception. In the first, Holigraph problem, G–I groups actually performed at, or slightly above, the level predicted by the model.

Individuals Versus Groups

These results clearly add support to the view that groups tend to make better decisions than individuals (Davis, 1969, p. 38). It is significant to find that these results obtain in decision situations which are similar to those encountered in business (unlike pure gambling and many of the other tasks used previously). While, with one exception, real groups did not perform as well as the best individual, G–I groups did better the performance of plurality vote synthetic groups. Previous studies (Davis, 1969) suggest that best-man types of pooling rules, such as the Lorge and Solomon's (1955) model, overestimate the quality of decisions real groups will achieve. Real groups seem to be generally

TABLE 4

PREDICTED AND OBSERVED PROPORTIONS OF OPTIMAL DECISIONS

Problem	P	Actual decisions		Synthetic decisions		
		I–G groups	G–I groups	R[a]	Best man	Plurality[b]
Holigraph (Problem 1)	.168	.316	.667	.601	.580	.316
Zirolex I (Problem 2)	.158	.158	.333	.577	.580	.000
Zirolex II (Problem 3)	.179	.211	.278	.627	.580	.111

[a] $R = 1 - (1 - P)^5$ where P = Proportion of I–G individuals making an optimal decision on a problem and R = Predicted proportion of groups making an optimal decision on a problem (see Lorge & Solomon, 1955).

[b] In only one case did a clear plurality for the optimal decision exist. In other cases, ties were resolved in favor of the optimal decision in order to provide the strictest possible test.

incapable of fully using the contributions of individuals who have high-quality decisions. However, groups do seem to benefit from deliberations in the sense that they make better decisions than would be obtained by pooling individual decisions using a plurality vote rule.

Of course, it is not completely reasonable to accept the best-man and Model A approaches to accurately predict group performance when the solutions to the problems are not of the "self-confirming" or eureka type. In tasks like economic decisions, the relative quality of alternative solutions may not be apparent to even the best individual problem solver, much less to the entire group. Thus the common assumption in both the best-man and Model A approaches, that if one person in the group finds the optimal solution it inevitably is adopted by the group, is an extremely strong assumption for the type of decision task used in many problem-solving studies, as well as in many real-life problems. Consequently neither the best-man nor Model A pooling rules constitute appropriate explanations of the difference between individual and group decision-making performance for noneureka tasks. Adequate explanation requires both an analysis of how any given persons obtain an optimal solution and how that solution gains group acceptance.

In the Holigraph task, comparisons of I–G individuals with both I–G groups and with G–I groups imply that both sets of groups were more conservative than were individuals. This result may be due to the description of the problem which indicated that the firm's board of directors tended to value caution. In the Zirolex I task groups were significantly more risky (in the variance sense) than were individuals. In that situation subjects were provided with no inference as to the organization's attitude toward risk taking, but they could infer that they should be risk neutral.

If the cultural values interpretation may be properly applied in this case, groups do appear to be more responsive than were individuals to the orientation toward risk if this orientation facilitates reaching high-quality decisions.

I–G Versus G–I Groups

One of the most surprising findings is that the inexperienced (G–I) groups made better decisions than did experienced (I–G) groups. In the absence of previous studies of this issue some speculation as to the reasons for the relative superiority of inexperienced groups is warranted. A likely reason would be the fact that I–G group members had already arrived at a decision and made a sort of "public stand" on that decision by handing it to the experimenter. Having taken a stand with respect to that decision, they become committed to it and were reluctant to change it (Kiesler, 1971). A similar explanation could account for Campbell's (1968) finding of the inferiority of group as compared to individual decisions. The effect of commitment appeared strongest in the case of G–I individuals who seldom deviated from group decisions. This finding supports the commitment interpretation since subjects were making their group decisions known to all members of the groups as well as to the experimenter.

Some evidence consistent with the commitment explanation was obtained from confidence measures on individual decisions. In every decision, individuals who selected the worst alternative were just as confident (differences were materially and statistically insignificant) of their decision as were subjects selecting the best and second-best alternatives. Since the majority of subjects in all three decisions made nonoptimal individual decisions in the I–G condition, their confidence in their nonoptimal individual positions probably forced a greater expenditure of additional time and more compromises (of a nonoptimal nature) in I–G groups than was required in the G–I groups where individuals had no opportunity to develop prior strong confidence in their own choices.

Another factor which could explain the relative inferiority of I–G compared to G–I group decisions involves group structure. One would expect that I–G groups would develop clearer definitions of structure than would G–I groups, since fewer group resources are required to understand the decision situation. But do clearer structures facilitate better decisions? One might expect the answer to be yes for I–G groups, since the resolution of divergent opinion is a major problem and some structure should facilitate the resolution of conflicts. However, for G–I groups the resources devoted to clarifying structure detract from those devoted to problem definition, evaluation, and other necessary elements of the decision-making process. Thus, one might expect that for I–G groups structural clarity and decision quality would be positively related, but for G–I groups the relationship would be reversed.

One important finding remains unexplained. In the Holigraph decision, G–I groups performed better than any of the synthetic decision rules. It would appear that certain kinds of decision theoretic problems are particularly amenable to group decisions. The Holigraph problem fits into this category, the Zirolex problem does not. Which features of the two problems determine this bifurcation remain open to conjecture.

If the findings of group problem-solving literature are to have any relevance to real organizations and decisions, it appears necessary to move experiments from the laboratory to the real world. This experiment may be regarded as transitional in nature. Business school subjects were used and economic decisions were involved. If these results can be extended to continuing groups, this study indicates that amount of prior exposure and/or commitment to decision alternatives, clarity of group structure, and the risk orientation of the organization may affect the relative quality of decisions of individuals and different types of groups. Verification of these findings with actual managers and continuing executive groups is desirable.

REFERENCES

Barron, F., & Mackenzie, K. *A constrained optimization model of risky decisions.* (Working Paper 41) University of Kansas, March 1971.

Bem, D. J., Wallach, M. A., & Kogan, N. Group decision making under risk of aversive consequences. *Journal of Personality and Social Psychology,* 1965, 1, 453–460.

Campbell, J. P. Individual versus group problem solving in an industrial sample. *Journal of Applied Psychology,* 1968, **52**, 205–210.

Cohen, J., & Hensel, M. *Risk and gambling.* New York: Philosophical Library, 1956.

Davis, J. *Group performance.* Reading, Mass.: Addison-Wesley, 1969.

Deets, M., & Hoyt, G. Variance preferences and variance shifts in group investment decisions. *Organizational Behavior and Human Performance,* 1970, **5**, 378–386.

Dion, K., Miller, N., & Baron, R. Group risk taking: A review of the risky shift phenomenon. In L. Berkowitz (Ed.), *Advances in experimental and social psychology.* Vol. 5. New York: Academic Press, 1972.

Dunnette, M. D., Campbell, J. P., & Jaastad, K. The effect of group participation on brainstorming effectiveness for two industrial samples. *Journal of Applied Psychology,* 1963, **47**, 30–47.

Green, P., Halbert, M., & Fitzroy, P. *Experiments on the value of information in simulated marketing environments.* Boston: Allyn & Bacon, 1967.

Hall, E. J., Mouton, J. S., & Blake, R. R. Group problem solving effectiveness under conditions of pooling vs. interaction. *Journal of Social Psychology,* 1963, **59**, 147–157.

Kiesler, C. *The psychology of commitment.* New York: Academic Press, 1971.

Kogan, N., & Wallach, M. A. Risk taking: A study in cognition and personality. New York: Holt, Rinehart & Winston, 1964.

Krantz, D., Luce, R., Suppes, P., & Tversky, A. *Foundations of measurement.* Vol. 1. New York: Academic Press, 1971.

Lorge, I., Fox, D., Davitz, J., & Brenner, M. A survey of studies contrasting the quality of groups performance and individual performance. *Psychological Bulletin,* 1958, **53**, 337–372.

Lorge, I., & Solomon, N. Two models of group behavior in the solution of Eureka-type problems. *Psychometrika,* 1955, **20**, 139–188.

Markowitz, H. *Portfolio selection.* New York: Wiley, 1959.

Siegel, S. *Non-parametric statistics.* New York: McGraw-Hill, 1956.

Simon, H. *The new science of management decision.* New York: Harper, 1960.

Slovic, P. Convergent validation of risk taking measures. *Journal of Abnormal and Social Psychology,* 1962, **65**, 68–71.

Stoner, J. A. F. A comparison of individual and group decisions involving risk. Unpublished master's thesis, Massachusetts Institute of Technology, School of Industrial Management, 1961.

Taylor, D. W., Berry, P. C., & Block, C. H. Does group participation when using brainstorming facilitate or inhibit creative thinking? *Administrative Science Quarterly,* 1958, 3, 23–47.

Tuckman, J., & Lorge, I. Individual ability as a function of group superiority. *Human Relations,* 1962, **15**, 45–51.

(Received April 3, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 4, 433-436

BRAINSTORMING BY INDIVIDUALS, COACTING AND INTERACTING GROUPS[1]

WARREN R. STREET[2]

Central Washington State College

Thirty-six 3-person groups of undergraduate volunteers worked in three different settings to produce answers to three brainstorming problems. The subjects worked independently in separate rooms (individual condition), independently in the same room (coaction condition), and cooperatively in the same room (interaction condition). Conflicting theories predicted lower productivity either in the interaction and coaction conditions, where co-workers were physically present (social facilitation theory), or only in the interaction condition, where group norms could be formed. Results showed that the independent and coaction conditions produced the same number of solutions and the interaction condition was significantly ($p < .001$) inferior to both. This suggests that public action, not physical proximity, is responsible for individual–group differences in brainstorming.

Several experiments performed over the last 15 years have demonstrated that individuals are superior to groups in providing answers to creativity problems requiring divergent thinking (Bouchard, 1972b; Bouchard & Hare, 1970; Campbell, 1968; Dunnette, Campbell, & Jaastad, 1963; Meadow, Parnes, & Reese, 1959; Rotter & Portugal, 1969; Taylor, Berry, & Block, 1958). There are obvious applications of these results, but their generalizability would be enhanced by an understanding of the causes for a given individual's decreased responsiveness in a group setting.

Researchers in brainstorming have suggested a number of reactions to the group setting that may take place within the individual member and result in decreased productivity. One possible explanation is that the individual member perceives the group's norm to be one of noninvolvement in the problem at hand. Each person waits for someone else to take the lead in producing solutions, and the result is low productivity for the whole group. Individual passivity in group settings has been found in research dealing with bystander intervention in emergencies (Darley & Latané, 1968; Latané & Darley, 1968; Latané & Rodin, 1969). If the same norm of noninvolvement affects brainstorming groups, training techniques or instructions which stress involvement and individual participation should increase productivity. Two recent studies (Bouchard, 1972a, 1972b) demonstrate that such techniques are effective in partially overcoming the inferior productivity of brainstorming groups.

A second, related explanation offered in brainstorming studies is that fear of social disapproval causes group members to censor their contributions, regardless of how much the experimenter encourages them to report every answer that comes to mind. This same fear of disapproval may also cause group members to perseverate on one class of similar, "safe" solutions and avoid plowing new ground in the search for solutions. Because it perseverates, a group is not likely to produce the diversity of solutions that would be obtained from the combined output of the members working independently. It follows that diversity of group-produced solutions could be increased by simply increasing the number of groups working on a problem. Thus, if the perseveration hypothesis is correct, the pooled output of many groups should be at least equal to the output of an equal number of isolates. Bouchard, Drauden, and Barsaloux's (1974) reexamination of data from past studies, however, shows that groups are still inferior to individuals when a pooling procedure is followed. While perseveration does not appear to account for group–individual differences, there are still other alternatives worthy of consideration.

---

[1] A version of this article was presented at the Western Psychological Association, Anaheim, 1973.

[2] Requests for reprints should be sent to Warren R. Street, Department of Psychology, Central Washington State College, Ellensburg, Washington 98926.

One such explanation for the inferiority of groups is offered by an application of Zajonc's (1965) theory of social facilitation. Zajonc contends that the presence of others leads to heightened arousal, or drive. This, in turn, causes an increased likelihood of giving well-learned responses to stimuli and decreased likelihood of giving poorly learned or unusual responses. Since these unusual responses are the ones called for in a problem requiring many creative answers, the presence of others should decrease the number of creative solutions. This, of course, is the observed result of brainstorming studies.

It is important to note that decrements in creativity, according to social facilitation theory, should result from the mere presence of others, regardless of whether they are interacting or simply working independently in close physical proximity (coacting). The brainstorming productivity of both interacting and coacting groups should be inferior to individual output if the social facilitation model is correct. The present study compared the productivity of individuals, coacting groups, and interacting groups in order to (*a*) replicate the group–individual differences found in earlier studies and (*b*) to examine the productivity of coacting groups where others were present but working independently.

METHOD

Subjects

The subjects were 108 undergraduate psychology students who volunteered to participate in the study. They formed 36 3-person groups which were homogeneous with respect to sex and of the same sex as the experimenter. The experimenters were 24 undergraduate students of experimental social psychology. The study was described to the experimenters as an exploratory study with no specific hypothetical outcome.

Design

The experiment was a repeated-measures design wherein each group of three subjects gave solutions for three different problems. Each problem was presented in a different setting. The order of presentation of settings, order of presentation of problems, and the assignment of problems to settings were all completely counterbalanced, resulting in 36 different combinations.

Procedure

Settings. Individual, interaction, and coaction settings formed the three levels of the independent variable of interest. In the individual setting, subjects were each placed in isolated sound-controlled cubicles furnished with a table, chair, pen, and paper. They had 10 minutes in which to write as many solutions to their problem as they could think of. In the interaction setting, subjects were all seated in the same room around one large table and were to produce solutions by talking freely and working together for 10 minutes. In the coaction setting, the subjects were all seated at individual desks in the same room, separated by about two feet. They were instructed not to communicate while they wrote their answers in the 10-minute time allotted. In the interaction setting, one subject, a volunteer, recorded the group's answers.

This general procedure differs from many past studies in one regard. In this study, the subjects recorded their answers by writing, while most previous studies have tape-recorded oral responses. In a study where written responses were used, Bouchard (1969) points out that a consistent response mode across treatment conditions is an important requirement for valid comparisons. In the present study, the coaction setting necessitated written, instead of oral, responding, and this mode of response was used in all settings.

Instructions. After arriving at a waiting room, each group was met by an experimenter who introduced himself, thanked them for coming, and explained the need to read instructions from a prepared text. The instructions for each setting were as follows:

> (Individual) Each of you is in a separate room. I can talk to all of you through this intercom. In a moment, I will give you a problem for which there are many creative answers. Your job is to come up with as many solutions to this problem as possible. Don't worry about whether your answers are realistic or practical. Just try to produce as many answers as possible. Write each answer down on the sheet in front of you. You will have ten minutes in which to work on this problem. Now, if you have any questions before I read the problem, just speak into the intercom and I'll answer your questions.
>
> (Interaction) We've placed you all in the same room and we want you to work on this task as a group. This means that you should talk freely to each other and work together to do as well as possible. In a moment, I will give you a problem, etc. . . . . (the remainder is identical to the Individual setting instructions).
>
> (Coaction) We've placed you all in the same room, but we want you to work on this task as individuals. This means that you should not talk to each other or look at each other's work.

We've given you each a separate desk so that you will work independently. In a moment, I will give you a problem, etc. . . . . (the remainder is identical to the Individual setting instructions).

Problems. The problems were similar to those used in other brainstorming studies and read as follows:

1. Height problem: We know today that the average height and weight of people is increasing. Suppose discoveries in medicine and nutrition accelerate this trend and that, 20 years from now, the average American is seven feet tall and that the average weight is 250 pounds. What would be the consequences? What adjustments would this situation require?

2. Thumbs problem: This is not likely to happen, but imagine for a moment what would happen if everyone born after 1980 had an extra thumb on each hand. This extra thumb would be built just like the present one is, but would be located on the other side of the hand. It would bend inward toward the fingers, just as the present one does. Here is the question: What benefits or difficulties will arise when people start having this extra thumb?

3. Tourists problem: Each year a great many American tourists go to visit Europe. But now suppose that our country wanted to get many more Europeans to come to America during their vacations. What steps can you suggest that would get many more Europeans to come to this country as tourists?

RESULTS

The dependent variable, group productivity, was defined as the number of nonoverlapping solutions produced by a 3-man group on a single problem. While subjects in the coaction and individual settings did not actually produce solutions as a group, their ideas were combined and scored as one group's output. These are typically called "nominal groups" in the literature. While the tourists problem appears to be of a qualitatively different nature than the other two problems, there were no significant differences in the mean numbers of answers given for different problems.

The data were analyzed by means of a 3×6 repeated-measures analysis of variance (ANOVA), testing for differences between three settings and six possible orders of settings. Since some previous studies report increased output following a "warm-up" problem in isolation, it was possible that setting orders may have been related to productivity. The mean productivities per group (Table 1) in the individual and coaction settings are almost identical and far superior to productivity in the interaction setting. The ANOVA shows the effects of settings to be significant beyond the .001 level ($F = 12.85$, $df = 2/10$), and Tukey's test of honestly-significant-differences shows no differences between the individual and coaction settings, both of which are higher ($p < .01$) in productivity than are groups in the interaction setting. There are no effects related to orders of settings ($F = 1.03$, $df = 5/30$, *ns*).

TABLE 1

MEAN NUMBER OF RESPONSES TO CREATIVITY PROBLEMS

Setting	Mean
Individual	27.53
Coaction	27.06
Interaction	17.72

DISCUSSION

In replication of similar studies, interacting groups were found to be inferior to individuals in the production of answers to problems of divergent thinking. The results of the present study go on to show that some factor other than the mere presence of two other workers is necessary to produce the inhibiting effect of groups. It is possible that larger numbers of co-workers could hamper productivity, but the present study indicates that working in the same room with others does not hamper creativity. It is only when one's actions become public knowledge that decreased performance results. This conclusion coincides with that offered by Cottrell (1972) in a recent review of social facilitation literature.

A methodological issue in the present study should be mentioned in this report. In the coaction and individual settings, all three subjects could record answers. To promote verbal exchange in the interaction setting, only one subject, the volunteer recorder, was given writing materials. This may have reduced the number of answers recorded in the interaction groups. In a brief follow-up study, 24 3-man groups solved one of the three original problems in the interaction setting. Each subject was given writing materials, and the group was given 20 minutes in which to work. Even under these favorable conditions, mean productivity was 22.83 solutions, still significantly less than individual output ($t = 1.70$, $p < .05$), although improved over the original interaction groups ($t = 1.78$, $p < .05$). The pattern of group-individual differences was unaltered by this change in method.

It seems likely that differences between group and individual output do not stem from physiological arousal but rather from compliance with group norms of noninvolvement. In applied settings, steps may be taken to compensate for these inhibiting norms. The experimental literature indicates that special motivational, training, and problem-solving techniques occasionally result in groups that are as creative as individuals. Rarely, however, do these groups surpass individual productivity. Time-consuming and expensive techniques to bolster group creativity do not appear to pay their way in numbers of ideas.

In deciding upon a problem-solving tactic, group techniques should not be rejected without consideration, despite the results of this and other brainstorming studies. These studies typically employ problems requiring the largest possible number of solutions, but many practical problems call for the production of a single, best solution. A combination of individual, anonymous production of a large number of alternatives and group deliberations on the best alternative may be useful in these circumstances. In addition, situations where relevant information is distributed among many people may justify a group problem-solving approach (Bouchard, 1971).

Where creativity and diversity are desired, however, individuals working alone should be most productive. The present study shows that the physical proximity of individual brainstormers is inconsequential, but that privacy or anonymity of response is a major factor contributing to productivity.

REFERENCES

Bouchard, T. J. Personality, problem-solving procedure, and performance in small groups. *Journal of Applied Psychology Moongraph*, 1969, **53**, (1, Pt. 2).

Bouchard, T. J. Whatever happened to brainstorming. *Journal of Creative Behavior*, 1971, **5**, 182-189.

Bouchard, T. J. A comparison of two group brainstorming procedures. *Journal of Applied Psychology*, 1972, **56**, 418–421. (a)

Bouchard, T. J. Training, motivation, and personality as determinants of the effectiveness of brainstorming groups and individuals. *Journal of Applied Psychology*, 1972, **56**, 324–331. (b)

Bouchard, T. J., Drauden, G., & Barsaloux, J. A comparison of individual, subgroup, and total group methods of problem solving. *Journal of Applied Psychology*, 1974, **59**, 226–227.

Bouchard, T. J., & Hare, M. Size, performance, and potential in brainstorming groups. *Journal of Applied Psychology*, 1970, **54**, 51–55.

Campbell, J. P. Individual versus group problem solving in an industrial sample. *Journal of Applied Psychology*, 1968, **52**, 205–210.

Cottrell, N. B. Social facilitation. In C. G. McClintock (Ed.), *Experimental social psychology*. New York: Holt, Rinehart & Winston, 1972.

Darley, J. M., & Latané, B. Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, 1968, **8**, 377–383.

Dunnette, M. D., Campbell, J., & Jaastad, K. The effect of group participation on brainstorming effectiveness for two industrial samples. *Journal of Applied Psychology*, 1963, **47**, 30–37.

Latané, B., & Darley, J. M. Group inhibition of bystander intervention in emergencies. *Journal of Personality and Social Psychology*, 1968, **10**, 215–221.

Latané, B., & Rodin, J. A lady in distress: Inhibiting effects of friends and strangers on bystander intervention. *Journal of Experimental Social Psychology*, 1969, **5**, 189–202.

Meadow, A., Parnes, S. J., & Reese, H. Influence of brainstorming instructions and problems sequence on a creative problem-solving test. *Journal of Applied Psychology*, 1959, **43**, 413–416.

Rotter, G. S., & Portugal, S. M. Group and individual effects in problem solving. *Journal of Applied Psychology*, 1969, **53**, 338–341.

Taylor, D. W., Berry, P. C., & Block, C. H. Does group participation when using brainstorming facilitate or inhibit creative thinking? *Administrative Science Quarterly*, 1958, **3**, 23–47.

Zajonc, R. B. Social facilitation. *Science*, 1965, **149**, 269–274.

(Received August 6, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 4, 437–444

CUE UTILIZATION AND EXPERT JUDGMENTS:
A COMPARISON OF INDEPENDENT AUDITORS WITH OTHER JUDGES

ROBERT H. ASHTON[1]

University of Texas at Austin

Cue utilization by independent auditors, a type of expert judge, was examined within an analysis of variance framework. After reading a brief narrative describing a hypothetical manufacturing company, 63 auditors judged the strength of payroll internal control as a function of 32 different combinations of six internal control indicators. The results demonstrated the overwhelming importance of main effects as compared to configural cue utilization. The average correlation for interjudge consistency was .70, while the average intrajudge correlation was .81. The average insight index (the correlation between an auditor's subjective weights and the statistical weights of the six cues) was .89. Similarities and differences between the present findings and those of studies involving other types of expert judges are discussed.

Expert judges are faced with the task of interpreting and integrating numerous items of information into an overall judgment or decision. To perform this task, a judge must weight items of information differentially and must qualify his interpretation of a given item when other considerations make such qualification necessary (Slovic, 1969). Because of the importance of such judgments in psychology, business, and elsewhere, researchers have recently conducted systematic studies in an effort to understand the manner in which experts use information. An important goal of such studies has been the *description* of expert judgments. If such judgments, and the information-processing activities which precede them, can be adequately described, then it may be possible to develop techniques to improve upon human judgment in various applied areas.

Initially, multiple regression techniques were used in an effort to identify both linear and configural cue utilization. Hoffman's (1968) discussion concerning the lack of success on the part of judgment researchers in identifying configural cue utilization with multiple regression models has been followed by a small number of studies conducted in an analysis of variance framework. A typical finding of such studies has been the overwhelming importance of main effects (representing linear cue utilization) as compared to interactions (representing configural cue utilization). This research has involved various types of judges and judgment tasks; for example, Veterans Administration (VA) hospital personnel (Rorer, Hoffman, Dickman, & Slovic, 1967), radiologists (Hoffman, Slovic, & Rorer, 1968), stockbrokers (Slovic, 1969; Slovic, Fleissner, & Bauman, 1972), and graduate students (Slovic et al., 1972).

The present article reports a study involving a different type of expert judge—independent auditors employed by public accounting firms. In the process of performing an audit, an independent auditor must frequently rely upon his professional judgment in combining various items of information into an overall evaluation. In addition to studying cue utilization via the analysis of variance model, the present study assessed (*a*) judgment consensus, or agreement across auditors at the same point in time using the same data; (*b*) judgment stability, or consistency over time for the same auditor using the same data (Goldberg & Werts, 1966); and (*c*) judgment insight, or the extent of agreement between the statistical description of the auditor's judgments provided by the analysis of variance model and a subjective description of those judgments provided by the auditor himself (Hoffman, 1960).

---

[1] Requests for reprints should be sent to Robert H. Ashton, Business–Economics Bldg. 324, University of Texas, Austin, Texas 78712.

METHOD

Subjects

The subjects were 63 practicing auditors who were employed by four public accounting firms in the Minneapolis/St. Paul area. The four firms consisted of two large national firms, a regional firm, and a local firm. The great majority of the auditors had 2 or 3 years of experience. A partner or manager in each firm was contacted to determine how closely the experience levels of the subjects coincided with the experience levels of those auditors who, in actual audits, make the type of judgment studied in the experiment. All firms indicated that the typical amount of experience required before an auditor is allowed to make this type of judgment is 2½ to 3 years. Therefore, most of the subjects were approaching the point in their careers at which they would actually make this type of judgment.

Task

The experimental task selected concerns the auditor's evaluation of his client's system of "internal control," which has been described as

> . . . the plan of organization and all of the coordinate methods and measures adopted within a business to safeguard its assets, check the accuracy and reliability of its accounting data, promote operational efficiency, and encourage adherence to prescribed managerial policies [American Institute of Certified Public Accountants, 1949, p. 6].

The selection, timing, and extent of auditing procedures applied in a particular engagement depend, in part, upon the judged effectiveness of existing internal control (Broeker, 1967; Mautz, 1964). It is important, therefore, that the process of judgment formation in the evaluation of internal control be subjected to systematic research.[2]

In the evaluation of a client's system of internal controls, auditors typically concentrate upon individual internal control subsystems; for example, cash receipts, cash disbursements, inventories, etc. The subsystem chosen in the present research was that of payroll, a subsystem of considerable importance because of the factor of relative risk. The characteristics of adequate payroll internal control should be familiar to all auditors. For the purpose of this experiment, six indicators of payroll internal control were selected. They were presented to the auditors in the form of questions. Each question was answered by the experimenter as "yes" or "no," and a "no" answer always represented a weakness in the internal control subsystem. It is important to recognize that the most common instrument used by auditors for reviewing internal control subsystems is the internal control questionnaire, which corresponds very closely to the stimulus presentations used in this experiment. Of course, the typical internal control questionnaire contains more than six questions.

Each subject was given a brief narrative describing a hypothetical manufacturing company. Included in this narrative was a description of some basic structural aspects of the company's payroll internal control subsystem. In conjunction with this unchanging background information, each subject judged the strength of internal control as a function of 32 different combinations of yes and no answers to the six questions. Their judgments were made on a scale from 1 (extremely weak) to 6 (adequate to strong). One of the stimulus presentations (hereafter called a "case") used in the experiment appears in Figure 1.

Each of the first five questions represents an aspect of payroll internal control which is recognized to be very desirable from the standpoint of sound internal control. The sixth question is included because auditors may use past results as a starting point for the examination of the current year; indeed, they may (perhaps erroneously) consider the past to be the best indicator of present conditions. Another reason that past results are important is that all aspects of an internal control subsystem may not be examined every year. Instead, only a few aspects may be examined each year and that information may be combined with the results of previous examinations in evaluating the effectiveness of the controls.

Design

Since six questions were employed, each of which could assume two answers, 64 possible patterns of answers (2^6) to the six questions were available. In order to enable the experimental task to be completed within a reasonable time period, only 32 patterns were actually used. Specifically, the 32 patterns were chosen in accordance with a one-half fractional replication of a 2^6 factorial design (Cochran & Cox, 1957). The estimation of four-factor and five-factor interactions is impossible with this particular design because their mean squares are included in the mean squares of main effects and two-factor interactions. This does not present a significant problem; much empirical evidence exists which indicates that such higher order interactions in studies of human judgment are negligible (Goldberg, 1968; Hoffman, 1968; Rorer et al., 1967). Since the data were analyzed separately for each auditor, there was no within-cell variance from which to construct the error term for the analysis; therefore, the mean squares from the three-factor interactions were combined to form the error term (Davies, 1956). The inability to evaluate three-factor and higher order interactions is not a significant disadvantage; a principle aim of the study was to assess the relative importance of main effects (repre-

---

[2] For a more complete discussion of (*a*) the importance of internal control evaluation in the performance of an audit and (*b*) the necessity of relying upon judgment in the evaluation of internal control, see Ashton (1973).

CASE NO. ______	YES	NO
1. Are the tasks of both timekeeping and payment of employees adequately separated from the task of payroll preparation?	X	
2. Are the tasks of both payroll preparation and payment of employees adequately separated from the task of payroll bank account reconciliation?		X
3. Are the names on the payroll checked periodically against the active employee file of the personnel department?		X
4. Are formal procedures established for changing names on the payroll, pay rates and deductions?	X	
5. Is the payroll audited periodically by internal auditors?	X	
6. Was the internal control over payroll found to be satisfactory during the previous audit?	X	

extremely weak	very weak	substantial weakness	some weakness	not quite adequate	adequate to strong
1	2	3	4	5	6

FIGURE 1. Example of an internal control case used in the experiment.

senting linearity) and two-factor interactions (representing configurality).

Analysis

Descriptive models of judgments. Analysis of variance was applied to the data collected from each auditor. In this context, a significant main effect for a particular question implies that the auditor's responses varied systematically with the value of that particular question. In other words, it made a difference to the auditor whether the question was answered yes or no. A significant two-factor interaction implies that the auditor was responding to particular patterns of answers to the questions; that is, the effect of variation of one question upon his responses differed as a function of the level of another question (Hoffman et al., 1968). After F ratios were obtained for the six main effects and 15 two-factor interactions, omega-square statistics (ω^2) were computed (Hays, 1963). In this context, omega-square is an estimate of the proportion of variance in judgment that is accounted for by the specific main effect or two-factor interaction. Therefore, omega-square is used to determine the extent to which each auditor used each of the six internal control questions (and their interactions) in formulating his internal control judgments.

Judgment insight. To assess the relationship between the statistical description of the auditors' judgments and a subjective description of those judgments, both types of descriptions were obtained. The auditor's subjective cue-weighting scheme was identified by having him distribute 100 points across the six questions in such a manner as to reflect the relative importance of each question to his judgments (Hoffman, 1960). Each auditor distributed his 100 points after having rated the 32 cases.

Order of presentation of the cases. The 32 cases were presented to the auditors in three different orders, with approximately one third of the auditors receiving cases arranged in each order. In Order 1, the first half of the cases was primarily composed of those in which only two questions were answered yes, while the second half was primarily composed of cases in which four questions were answered yes. This pattern was reversed in Order 2, while in the third order, a "four-yes" case always followed a "two-yes" case. The effect of order of presentation upon both statistical weights and subjective weights was assessed.

Judgment consistency. The two types of judgment consistency that were investigated in the study were (*a*) consensus, or consistency across auditors at the same point in time using the same data, and (*b*) stability, or consistency over time for the same auditor using the same data. Estimates of both types of consistency were expressed as correlation coefficients.

Procedure

The experiment was conducted by visiting the public accounting firms and meeting with the auditors in small groups. The participants were given a brief oral introduction to the judgment task and were then presented with the case materials. The auditors typically worked at the task for 30 to 40 minutes. The second administration of the experi-

TABLE 1

RELATIVE IMPORTANCE OF MAIN EFFECTS AND TWO-FACTOR INTERACTIONS TO THE AUDITORS

Main effect or interaction	No. auditors for whom significant*	Average ω^2	
		Significant ones only	All auditors ($N = 63$)
1	60	27.4	26.2
2	60	26.3	25.2
3	46	8.7	6.7
4	44	9.8	7.1
5	50	10.5	8.5
6	27	14.5	6.6
12	15	3.5	1.1
13	6	2.7	.4
14	8	2.3	.6
15	6	3.6	.5
16	7	1.9	.4
23	8	2.7	.6
24	8	3.2	.7
25	5	2.8	.4
26	7	1.5	.4
34	3	3.7	.4
35	1	1.8	.1
36	1	1.7	.2
45	2	2.2	.2
46	4	1.8	.2
56	1	.6	.3

* Significant at the .05 level. The great majority of significant main effects were significant well beyond the .01 level, but almost no interactions were significant at that level.

ment, which was necessary in order to assess consistency of judgment over time, was conducted in the same manner as the first administration.[3] The time interval between the two administrations ranged from 6 weeks to 13 weeks. Of course, the tactical procedures involved in making such judgments in actual auditing situations differ from those described in this study. For example, auditors do not make a series of similar judgments within such a brief time span, nor is a subsequent audit performed on the same client 6 to 13 weeks later. These simplifications were made necessary by practical considerations.

RESULTS

Main Effects and Interactions

An indication of the relative importance to the auditors of the main effects and interactions is presented in Table 1.[4] The percentage of variance in judgment accounted for by the six main effects combined ranged from 48% to 96%, with an average of 80.2%, while the range for the 15 two-factor interactions combined was 0% to 17%, with an average of 6.4%. (On the second administration, the average for main effects and interactions was 79.9% and 5.5%, respectively.) At least five of the six main effects were significant for two thirds of the auditors, while the number of interactions reaching significance was between zero and two for almost 80% of the auditors. On the average, the *least important* main effect (representing Question 6) accounted for six times as much variance in judgment as did the *most important* interaction (representing Questions 1 and 2). Apparently the auditors did not look for patterns of answers to the six questions; rather, they evaluated the effect of each question independently of the effects of other questions in the same case.

Individual Differences in Cue Utilization

A point of agreement between the results of this study and those of previous studies concerns the extent of individual differences in cue utilization. On the average, main effects representing Questions 1 and 2 accounted for a total of 51.4% of the total variance in judgment. This represents a sizable proportion of the 80.2% accounted for by the six main effects. Notice that both questions deal with separation of duties, an internal control characteristic which is crucial in the payroll subsystem because of the risk factor. The two questions were almost equally important; Question 1 accounted for 26.2%, and Question 2 accounted for 25.2% of the total variance. On the average, the other four questions were approximately equally important. Although Questions 1 and 2 were considered most important overall, such individual differences existed among the auditors that each question was considered most important by at least one auditor, and each question was considered least important by at least one auditor. Examples of individual differences in cue utilization are presented in Table 2.

---

[3] An exception was made in the case of the regional firm, in which it was necessary to conduct the second administration by mail. Five of those auditors did not respond; therefore, only 58 of the 63 auditors repeated the experiment.

[4] Unless otherwise specified, the data reported are taken from the first administration of the experiment. The results from the second administration were usually very similar and are presented where differences (or particular similarities) are of interest.

Judgment Insight

An insight index was computed for each auditor. This index took the form of a correlation between subjective weights and statistical weights across the six questions. Before computing the correlations, statistical weights were adjusted to include the effects of interactions (Slovic, 1969). It must be emphasized that the insight index was computed in order to provide a general indication of degree of relationship (agreement). One must be careful not to place too much confidence in the specific values of the insight indexes because correlations computed over such a small number of factors are likely to be highly unstable. In general, the agreement between subjective weights and statistical weights was quite strong. The average value of the insight index was .89; two thirds of the indexes were in the .80s and .90s, and only three were below .60.

Order of Presentation of the Cases

While the importance of total main effects and total interactions was not influenced by order of presentation, the importance of Questions 1 and 2 differed substantially. For those auditors who received Order 2, in which relatively strong cases of internal control appeared early in the sequence, Questions 1 and 2 were apparently less important than for those auditors who received the cases in a different sequence. The same effect occurred for subjective weights and statistical weights on both administrations of the experiment. Data are presented in Table 3. This implies that the importance of an internal control cue is partially determined by the context in which it appears. It also emphasizes, in another manner, the relatively strong agreement between the statistical weights attached to the questions by omega-square and the subjective weights assigned to them by the auditors.

TABLE 2

Examples of Differences in the Importance of the Questions to the Various Auditors

Most important question	Auditor no.	Percentage of total variance in judgment accounted for by question					
		1	2	3	4	5	6
1	24	79	2	1	9	1	*ns*[a]
1	44	74	7	4	1	1	*ns*
2	2	*ns*	67	*ns*	*ns*	9	7
2	21	5	67	13	*ns*	3	3
3	29	7	24	29	10	4	*ns*
3	37	16	12	33	*ns*	16	*ns*
4	27	14	7	12	28	*ns*	*ns*
4	57	3	14	14	33	14	*ns*
5	56	6	6	3	12	45	8
5	59	11	6	3	3	29	29
6	22	*ns*	2	2	1	18	66
6	42	*ns*	*ns*	3	*ns*	26	48

[a] Not significant at the .05 level.

TABLE 3

Effect of Order of Presentation of the Cases on the Combined Importance of Questions 1 and 2

Order	Average statistical weight		Average subjective weight	
	Administration 1	Administration 2	Administration 1	Administration 2
1	54.5	55.4	48.7	52.3
2	45.4	42.3	44.4	42.2
3	53.8	52.8	49.2	51.1

Judgment Consistency

Concerning judgment consensus, the correlations between the ratings of all pairs of auditors on the 32 cases were computed as an estimate of the extent of agreement between their internal control judgments. The average correlation was approximately .70. The range of correlation coefficients was generally from the lower .40s to the higher .80s, but a small number of coefficients were between .06 and .40 (about 3%) and between .90 and .93 (about 1%). Generally, the more discrepant were the cue-utilization schemes of two auditors, the lower was the correlation between their resultant judgments. Consensus differed very little as a function of firm affiliation or experience level.

Concerning judgment stability, which was assessed for each auditor by correlating his ratings from the two administrations of the experiment, the coefficients ranged from a low of .43 to a high of .96, with an average of .81. Only four of the 58 coefficients were below .70. Again, the coefficients differed very little by firm or experience level.

DISCUSSION AND CONCLUSIONS

The present study has revealed some similarities, and some differences, between judgments made by independent auditors and other types of expert judges. These similarities and differences are discussed below. Possible implications of the present findings for the auditing profession are discussed elsewhere (Ashton, 1973).

The most striking similarity concerns the overwhelming importance of main effects as compared to interactions when the cue-utilization strategies were represented by the analysis of variance model. In the present study, it is perhaps unrealistic to expect sizable configural cue utilization. Since a yes answer to *any* question represented a favorable aspect of internal control (and, likewise, a no answer always represented an unfavorable aspect), one would not expect different patterns of answers to reverse the implications of particular cues. On the other hand, if judgment processes of experts really include a sizeable configural component, the present study should have allowed the demonstration of that component to a greater extent than previous studies. The judgment task used in this study was simpler than those which have typically been used. The probability that the descriptive models which emerged from this study may represent the judges' simplification of an overly difficult task would appear to be lower than has been the case in previous studies.

The fact that very small configural effects were found, even in this relatively simple judgment task, reinforces even more strongly the results of previous studies. Hoffman et al. (1968) found that approximately 90% of the variance in judgments could be explained by main effects. The average percentage of judgment variance explained by interactions in the study by Slovic (1969) was 6% and by Slovic et al. (1972) was 4%. Even smaller percentages were reported by Rorer et al. (1967). In those studies, as in the present study, it appears that small configural effects may be identified when certain patterns augment or attenuate the importance of particular cues (Hoffman et al., 1968). This situation is sometimes called "ordinal interaction" as opposed to "disordinal interaction" (Digman, 1966).

Some differences between prior studies and the present study should be noted: first, the present study involved a larger number of subjects (63) than did any of the studies previously mentioned (their range was 2 to 24). While it is true that analysis of variance was applied to the data of each subject separately, the consistency of the analysis of variance results over a larger number of subjects lends additional credibility to the findings. Second, the format of the stimulus materials used in this study corresponded more closely to the "real world" than did formats used in some previous studies; for example, whereas stockbrokers do not, in actual practice, make evaluations based upon dichotomized variables—"up" versus "down," "good" versus "poor," "many" versus "few" (Slovic, 1969; Slovic et al., 1972)—auditors do evaluate internal control effectiveness via a questionnaire containing a series of yes and no answers.[5] Third, the variables of experience level and firm affiliation were given explicit attention in this study. The data were analyzed to determine any differences related to these variables. No significant differences concerning configural processing were found.

In this study, as in previous studies, judgment stability was greater than judgment consensus. When data for individual auditors were examined, it was found that stability, like consensus, was related to the use of particular internal control questions. Auditors who relied upon the same one or two questions to approximately the same extent on both testings exhibited relatively high stability. Those who changed their patterns of reliance exhibited relatively low stability.

---

[5] On actual questionnaires a third category of "not applicable" is sometimes included. It was not included in the experiment because only six questions were used and all were presumed relevant. As usual, it is difficult to generalize beyond the levels actually used in the experiment.

Other judgment studies of this type have not assessed stability over a significant period of time. Hoffman et al. (1968) had each judge make two judgments of each stimulus case, apparently separated by only a few minutes, and found that the judges' reliabilities ranged from .60 to [illegible] with a median of .80. In another study Hoffman (1960) found reliabilities in the mid-.80s. Length of time between judgments may be an important factor in judgment stability. For the present study, a slightly negative relationship between stability and length of time was found. As the time interval increased from 43 days to 94 days (in roughly 10-day intervals), average stability consistently decreased from .83 to .76.

It is also interesting to note that judgment consensus was greater in this study than in studies of other types of expert judges (see Goldberg, 1968). For example, in the Hoffman et al. (1968) study, the range of correlation coefficients was −.11 to .83, with a median of .38. In that study, approximately 60% of the coefficients were below .40.

The degree of judgment insight reported for other types of expert judges appears to be much lower than that reported in the present study (Hoffman, 1960; Slovic, 1969; Slovic et al., 1972). Only one specific average value for an insight index has been presented; it was .34 and was based upon the magnitude of effect index (Slovic et al., 1972). The same study found a negative relationship (Spearman $r = -.43$) between length of experience and degree of insight. In the present study, insight did not differ significantly as a function of experience. Of course, the relative simplicity of this particular judgment task may have contributed to the magnitude of insight indexes obtained.

The data collected in the present study cannot provide an explanation for the findings that independent auditors exhibited greater consensus and greater insight than have other types of judges. One may speculate, however, that the educational experiences of auditors —both in college and in preparation for the Certified Public Accountant examination—are at least partly responsible for these differences. The study of internal control in collegiate auditing courses is not likely to differ substantially among reputable colleges and universities. Characteristics of "good" internal control are defined quite clearly and are discussed in numerous publications with which the great majority of auditors should be very familiar (e.g., American Institute of Certified Public Accountants, 1949, 1963).

The extent to which the internal control evaluation situation is well defined may be contrasted with that pertaining to the medical diagnosis situation (Hoffman et al., 1968) and the stock evaluation situation (Slovic et al., 1972), in which the meanings of particular cues are likely to be much more ambiguous in relation to the variable to be judged. This may provide a somewhat more basic explanation for the observed differences between auditors and other types of expert judges.

The importance of individual judgment in accounting and auditing, and in business generally, should not be overlooked by judgment researchers. In business, as in other areas, the consequences of decisions which include individual judgment as a major input are often quite significant, and current techniques of judgment research are generally applicable.

REFERENCES

American Institute of Certified Public Accountants. *Internal control.* New York: AICPA, 1949.

American Institute of Certified Public Accountants. *Auditing standards and procedures.* New York: AICPA, 1963.

Ashton, R. H. Judgment formation in the evaluation of internal control: An application of Brunswik's lens model. Unpublished doctoral dissertation, University of Minnesota, 1973.

Broeker, M. M. Audit problems relating to review of internal control. *The Journal of Accountancy,* 1967, 123, 75–78.

Cochran, W. G., & Cox, G. M. *Experimental designs.* New York: Wiley, 1957.

Davies, O. L. (Ed.). *The design and analysis of industrial experiments.* New York: Hafner, 1956.

Digman, J. M. Interaction and non-linearity in multivariate experiment. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology.* Chicago: Rand McNally, 1966.

Goldberg, L. R. Simple models or simple processes? Some research on clinical judgments. *American Psychologist,* 1968, 23, 483–496.

Goldberg, L. R., & Werts, C. E. The reliability of clinicians' judgments: A multitrait-multimethod approach. *Journal of Consulting Psychology*, 1966, **30**, 199–206.

Hays, W. L. *Statistics*. New York: Holt, Rinehart & Winston, 1963.

Hoffman, P. J. The paramorphic representation of clinical judgment. *Psychological Bulletin*, 1960, **57**, 116–131.

Hoffman, P. J. Cue-consistency and configurality in human judgment. In B. Kleinmuntz (Ed.), *Formal representation of human judgment*. New York: Wiley, 1968.

Hoffman, P. J., Slovic, P., & Rorer, L. G. An analysis-of-variance model for the assessment of configural cue utilization in clinical judgment. *Psychological Bulletin*, 1968, **69**, 338–349.

Mautz, R. K. *Fundamentals of auditing*. New York: Wiley, 1964.

Rorer, L. G., Hoffman, P. J., Dickman, H. R., & Slovic, P. Configural judgments revealed. *Proceedings of the 75th Annual Convention of the American Psychological Association*, 1967, **2**, 95–96. (Summary)

Slovic, P. Analyzing the expert judge: A descriptive study of a stockbroker's decision processes. *Journal of Applied Psychology*, 1969, **53**, 255–263.

Slovic, P., Fleissner, D., & Bauman, W. S. Analyzing the use of information in investment decision making: A methodological proposal. *Journal of Business*, 1972, **45**, 283–301.

(Received July 20, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 4, 445–451

ROLE OF THE RATER IN PERFORMANCE APPRAISAL[1]

RICHARD J. KLIMOSKI[2] AND MANUEL LONDON

Ohio State University

Ratings for performance appraisal are frequently made by supervisors. In the present study, judgments of effectiveness for hospital nurses were obtained from the nurse herself and her peers in addition to her supervisor, using the same rating form. Factor analysis indicated that each rating source could be clearly identified and characterized. The data reaffirmed the notion that interrater disagreement may reflect systematic rater bias as well as meaningful differences in the ways in which judgments are made. Implications for comprehensive appraisals are suggested.

The problems, errors, and dilemmas inherent in ratings have been well documented (e.g., Barrett, 1966, Guion, 1965; Schmidt & Kaplan, 1971), and suggestions have been made for content changes, specific formats, and training (e.g., Kavanaugh, 1971). Similarly, numerous authors have been quite critical of current practice (Wallace, 1965), and some (Odiorne, 1964) have suggested the elimination of judgmental criteria entirely. The latter seems unlikely to occur for as long as there are people supervising others it appears there will be subjective (albeit covert) evaluations. Furthermore, there are those who agree with Blum and Naylor (1968) that subjective judgments are "a very rich source of data [p. 198]" that should not be abandoned but somehow gathered or transformed to meet its potential (see also Campbell et al., 1970, p. 123). Moreover, a number of these advocates appear to concentrate their efforts and optimism not on the development of specific rating forms per se, but instead on the judge, the rater, or as he sometimes is called, the observer.

"The ideal rater who observes and evaluates what is important and reports his judgment without bias or appreciable error does not exist, or if he does, we don't know how to separate him from his less effective colleagues [Barrett, 1966, p. 7]." While this may be true, Barrett feels he can continue and offer guidelines which, if followed, tend to maximize the potential accuracy of ratings. In doing so he places great stress on the importance of the organizational position of the rater relative to the person being rated, for "The position of the rater . . . determines, in part, the extent and nature of his opportunity to observe, the quality of his judgment and the appropriateness of his point of view [1966, p. 102]." The positions that he considers are those of self, superiors, peers (those who function on approximately the same organizational level as the person rated), subordinates (those who are supervised by the person rated), and the field review specialist (an outsider trained in performance measurement). And, after reviewing the strengths and liabilities of using ratings made by observers in these various positions, Barrett concludes that for certain purposes the less frequently used raters (e.g., peers or subordinates) appear to offer great benefits to an appraisal program. The focus of the present study is on judgments made by persons in the first three of the positions listed—supervisors, peers, and self.

By far, supervisors' judgments comprise the bulk of the ratings generated in industry for appraisals (cf. Campbell et al., 1970), and it is probably safe to say that most research on appraisals tends to be directed toward them (e.g., Barrett, 1966; Whisler & Harper, 1962). However, a growing body of literature on the potential usefulness of peer ratings exists that commands attention. Research on peer judgments, starting in the late forties (e.g., Hollander, 1954; Weitz, 1958; Wherry & Fryer, 1949) and continuing through the present (e.g., Amir, Kovarsky, & Sharan, 1970; Gordon & Medland, 1965; Mayfield, 1970, 1972) is impressive. Excellent reliabilities and, when used as predictors, good validities have been obtained by numerous authors in several types of organizations (Campbell et al., 1970, p. 112). In contrast, self-ratings have a less well-documented history and have been used in industry as an adjunct to other effectiveness data, primarily providing a vehicle for more effective appraisal interviews (Whisler & Harper, 1962). Current thinking is that no one position or organization vantage point can provide the information necessary to determine a person's effectiveness. Thus, recent literature suggests that one should gather ratings from several judges, at times from all three positions mentioned (Kavanaugh, MacKinney, & Wolens, 1971; Lawler, 1967). Indeed, this is a strategy in recent construct validity research. Yet despite this advocacy it is not clear why combinations of various ratings should be effective.

---

[1] The authors wish to thank Milton D. Hakel and Robert J. Wherry, Sr. for their helpful comments on an earlier draft of this article. Robert J. Wherry's advice on data analysis was invaluable.

[2] Requests for reprints should be sent to Richard J. Klimoski, Department of Psychology, College of Social and Behavioral Sciences, Ohio State University, 404-C West 17th Avenue, Columbus, Ohio 43210.

A few authors have offered some plausible arguments. For example, it appears that superiors and peers perceive different aspects of a person's behavior. Disagreement in this case should not be viewed as a mark of unreliability but instead as a potentially valid indicator that differing aspects of a worker's actions are being accurately perceived and reported (Norman & Goldberg, 1966). The personality of the rater (Schneider & Bayroff, 1953) and the rater's effectiveness on his own job (Kirchner & Reisberg, 1962) have been shown to differentially affect ratings. Guion (1965, pp. 472–473) has suggested that raters in different positions may, in fact, be using different percepts or dimensions in their evaluations of a worker. Thus, the additional validity obtained may be due to the more complete coverage of a true or "ultimate" criterion which, by nature, is complex. In particular, Guion feels that evaluations reflected in peer nominations may be fairly well differentiated (factorially complex) and suggests that "If so, peer nominations might well be subjected to dimension analysis for possibly greater predictive value [Guion, 1965, p. 473]." This is exactly what Mayfield (1972) did in his study. As Guion hypothesized, only one of the dimensions actually predicted Mayfield's criteria of survival and production of life insurance salesmen thus supporting this type of analysis.

In addition to increasing the sampling of the job behavior domain, combinations of various ratings should be more effective than single ratings by allowing the estimation of systematic error variance due to chance response tendencies and response sets. Such identifiable rater errors include central tendency, leniency, and faking discrimination by giving variable and inconsistent ratings (Dunnette, 1966). In Cronbach's (1970) terms, such temporary or lasting, specific variance can then be classified as random error. However, temporary, general variance still remains as systematic error. Systematic error variance due to changes in the ratee's performance and to the situation may be discovered by analyzing multiple ratings made at two or more intervals in time. In this way, temporary, general variance could be classified as random error allowing lasting, general variance to be isolated. Whether multiple ratings are made at more than one point in time or not, it is necessary to separate systematic rater bias from differences in the behavioral elements observed by the raters at their respective vantage points.

The present study addresses itself to the issue of dimensional complexity of performance appraisal judgments. Specifically, it examines ratings made at one point in time by supervisors, by peers, and by self-ratings in an attempt to discover both the nature and the complexity of these judgments. In doing so it was felt that the results of such an analysis could contribute to our understanding and usage of alternatives in building performance appraisal systems.

METHOD

The data were collected as part of a larger study concerned with the effects of possible moderators of the relationship between performance and satisfaction (London & Klimoski, 1973a, 1973b). One hundred and fifty-three registered nurses (RN) from four Columbus, Ohio area hospitals served as subjects. No attempt was made to split the sample according to hospital or functional unit (e.g., medical and surgical wards (pediatrics, obstetrics, etc.) because of the small number of subjects that would result in each category. All RNs working on the day shift in each hospital were required to fill out the questionnaires as part of a research study carried out in conjunction with the School of Nursing at the Ohio State University. In the study 202 RNs participated; however, 49 questionnaires were eliminated from the sample due to lack of self-ratings, supervisor ratings, and peer ratings for each subject.

The RNs filled out the research questionnaire during working hours in small groups. Head nurses completed their questionnaires at other times. The questionnaire of interest to the present study consisted of 19 measures of effectiveness. For each measure RNs were asked to simultaneously rate and rank their peers and themselves (head nurses had to rate and rank their subordinate RNs on the same items) on a 20-point continuum. The nurses first listed the names of the RNs in their units next to a letter and placed the letter corresponding to each name on the appropriate place on the scales. It was hoped that this method would encourage greater attention to scale content and, hence, greater care in making ratings.

The measures of effectiveness used in this questionnaire were obtained from performance appraisal forms used by [illegible] different hospitals including those sampled. The investigators tried to select characteristics (both behaviors and traits) which represented the range of effectiveness covered in the appraisal forms. The characteristics are listed in Table 2. The format of the rating scales used in the study was different from the appraisal forms used for administrative decisions in the hospitals; included were 18 characteristics of performance plus a measure of overall effectiveness. Instructions to the subjects emphasized that the questionnaires would be used for research purposes only and that complete anonymity would be maintained in recording the ratings. Although performance ratings used for research purposes are likely to be different from those used for pay raises or promotions (Lawler, 1967; Oberg, 1972), they may be less likely to be plagued by halo, and therefore underlying dimensions of effective performance will be more evident. Such an analysis may aid in constructing more valid performance appraisal forms.

Analyses

A 60 × 60 intercorrelation matrix of 20 variables (a total score of these ratings, 18 ratings on individual characteristics, and a single overall rating of effectiveness) for each of the three rater/sources was obtained. This matrix was then submitted to Wherry's (1959) hierarchical method of factor analysis. It is recognized that the number of subjects (153) is smaller, relative to the number of variables (60), than is usually considered desirable for a factor analysis by statisticians. However, since the aim is a practical one of understanding dimensions underlying the various ratings in order to determine which variables might be used for performance measures, the factor analysis should provide valuable knowledge beyond the intercorrelation matrix.

The hierarchical factor solution has been described by Wherry as follows:

> The Wherry hierarchical factor analysis program uses the highest correlation in a given row as communality estimates. (For less than 40 variables multiple R squares are used.) Factors are extracted by the principal-axis method until residuals are less than $1/\sqrt{N}$ or until the last eigenvalue is lower than .300. The Harman–Jones minimal residual method (Harman & Jones, 1966) is then used to stabilize the correct communalities, and the resulting corrected factor loadings are subjected to a varimax rotation.
>
> The actual hierarchical rotation is achieved by assigning variables to clusters corresponding to the varimax factors where items are assigned to the cluster on which they have the highest loading. The loadings are then used to create a theoretical R matrix with the corrected communalities in the diagonals, and the Thurstone multiple group method is used to obtain cluster intercorrelations. These are in turn factored, and the uniqueness for each cluster is obtained. A transformation matrix is obtained which transforms the original varimax loadings to a hierarchical pattern. All overlap now appears on the higher order factor, while the group factors now have simple structure when plotted against each other.[3]

RESULTS

A summary multitrait-multirater matrix with median intercorrelations for each triangle and validity diagonal is presented in Table 1. These data give an indication of the magnitude of the correlations in the different parts of the 60 × 60 matrix. It is apparent from the high-median intercorrelations in the heterotrait-monomethod triangles that strong rater-group specific bias differentiates self-ratings, supervisor ratings, and peer ratings from each other.

The results of the hierarchical factor analysis are presented in Table 2. One general and five subgenerals were extracted. (In this analysis original communality estimates ranged from .45 to .85, with a median of .71. Residuals ranged from +.149 to −.157.) This factor solution was selected by comparing a number of hierarchical solutions of the 60 × 60 intercorrelation matrix in which three through seven factors were extracted and by examining the results of individual

---

[3] Personal communication to the authors from Robert J. Wherry, Sr., October 17, 1973. See Wherry (1959) for a similar discussion.

TABLE 1

SUMMARY MULTITRAIT-MULTIRATER MATRIX WITH MEDIAN INTERCORRELATIONS

Measure	1	2	3
1. Self-ratings	.39**		
2. Supervisor ratings	.05 .03 .04	.56**	
3. Peer ratings	.08 .03 .07	.21** .18* .17*	.71**

Note. $N = 153$; ⬭ = validity diagonals; ◺ = heterotrait–heterorater triangle; ◺ = heterotrait–monorater triangle.
* $p < .05$.
** $p < .01$.

factor analyses performed on the 20 × 20 matrices of each rater group. (These data are not reported in this study.) In each factor analysis of the 60 × 60 matrix, three orthogonal factors representing self-ratings, supervisor ratings, and peer ratings emerged. Similarly, the separate factor analyses of the data of each rater group consistently resulted in a strong general factor made up of rater-group specific bias and interrater bias. (These two sources of error, however, are inseparable in the individual factor solutions.) This general factor accounted for more than 85% of the common variance leaving very low loadings on the subgenerals in all cases. At least one subgeneral factor emerged from every solution. The six factors extracted from the larger analysis combining the self-ratings, supervisor ratings, and peer ratings distinguishes between rater-group specific bias and interrater bias. It allows a comparison between the specific factors which emerged from the individual analyses. This indicates that factors interpretable in terms of traits or behaviors can be separated from rater halo.

It is evident from Table 2 that there is considerable halo common to supervisors and peers while strong rater specific biases also exist. Factors 1, 2, and 3 clearly indicate that self, supervisors, and peers each rate performance from a different perspective. Self-ratings account for 96% of the total variance of Factor 1; supervisor ratings account for 97% of the total variance of Factor 2; and peer ratings account for 99% of the total variance of Factor 3. Factor 1 accounts for 23% of the common variance of the factor analysis, while Factors 2 and 3 account for 15% and 29%, respectively.

The general factor (Factor A) accounts for 11% of the common variance. It indicates that supervisors and peers overlap in the way they make their ratings. Twenty supervisor ratings and 19 peer ratings load highly on this factor, while 19 self-ratings do not (a "significant" loading was arbitrarily taken to be greater than .20). It is interesting to note that the total scores for the self-ratings ($\bar{X} = 267.5$, $SD = 44.3$) were significantly higher than supervisor ratings ($\bar{X} = 246.7$, $SD = 60.0$, $t = 3.44$, $df = 304$, $p < .01$) and peer ratings ($\bar{X} = 248.6$, $SD = 50.5$, $t = 3.48$, $df = 304$, $p < .01$), whereas supervisor and peer ratings did not differ significantly. The general factor indicates that the significant differences are not merely differences in degree between self-ratings, peer ratings and supervisor ratings, but also a difference in dimensions evaluated by the raters. This could be interpreted as indicating that a form of "justifiable halo" exists insofar as two sets of raters responded to the same set

TABLE 2

HIERARCHICAL FACTOR ANALYSIS OF SELF-RATINGS, SUPERVISOR RATINGS, AND PEER RATINGS

Variable	A			1			2			3			4			5			h^2		
	General factor			Self-bias			Supervisor bias			Peer bias			Job competence			Effort					
	Self	Supervisor	Peers	Self	Supervisor	Peers	Self	Supervisor	Peers	Self	Supervisor	Peers	Self	Supervisor	Peers	Self	Supervisor	Peers	Self	Supervisor	Peers
1. Total Score	−01	*47*	*35*	*90*	−09	−02	−03	*68*	05	01	03	*92*	15	*53*	*25*	*24*	*24*	14	.89	1.00[a]	1.0
2. Technical competence	−01	*30*	*22*	*60*	−03	11	02	*41*	12	−07	15	*58*	*44*	*60*	*45*	−11	−06	−06	57	65	62
3. Ability to organize and schedule work loads	03	*29*	*20*	*64*	02	09	−04	*45*	03	02	10	*68*	*23*	*66*	*35*	06	−05	−04	47	74	64
10. Adaptability in emergencies	00	*28*	19	*65*	06	07	−03	*45*	06	−02	02	*64*	*28*	*58*	*41*	02	00	−11	50	62	63
20. Overall effectiveness	05	*34*	*24*	*70*	−13	−01	00	*56*	03	−01	01	*77*	*25*	*51*	*25*	09	17	01	56	74	71
4. Skill in planning	04	*31*	*23*	*61*	−02	−04	−03	*50*	07	02	06	*65*	19	*58*	*27*	09	06	00	42	69	55
12. Dependability	10	*34*	*26*	*68*	−03	−13	07	*53*	04	−02	−01	*76*	*33*	*52*	08	18	10	10	62	68	68
11. Quality of work	05	*35*	*37*	*58*	−07	−06	−01	*53*	06	−06	−04	*77*	13	*47*	19	13	14	05	38	65	78
9. Complete work schedule	06	*32*	*21*	*55*	02	04	−04	*46*	01	00	12	*73*	09	*50*	*21*	15	07	03	34	58	62
5. Acceptability of completed work	06	*35*	*23*	*51*	−10	03	−04	*51*	04	01	01	*68*	−01	*40*	16	19	15	09	30	58	55
6. Attendance and promptness	02	*38*	18	*56*	−09	−04	−09	*44*	−03	03	04	*65*	13	*32*	08	09	17	−06	35	48	47
19. Skill in communication	03	*35*	*24*	*64*	−16	09	−03	*49*	04	−06	−03	*65*	07	*25*	*22*	16	*27*	08	45	52	54
8. Amount of work	09	*37*	*21*	*64*	−11	−07	00	*57*	03	−02	−01	*69*	06	*44*	14	*22*	18	04	47	70	55
13. Willingness to perform	13	*36*	*28*	*64*	−12	00	−03	*54*	05	04	−02	*62*	−11	*27*	−04	*34*	19	*26*	56	55	53
14. Observe rules and regulations	05	*36*	*27*	*36*	−10	−14	02	*47*	07	−16	08	*58*	−06	16	−12	*21*	*29*	*27*	21	48	52
15. Effort applied	13	*37*	*30*	*62*	−14	04	01	*57*	04	04	−06	*73*	03	*28*	−02	*29*	*24*	*21*	49	62	67
17. Impress visitors	*21*	*41*	*24*	*44*	−07	06	14	*51*	−01	01	04	*63*	−02	14	−08	*36*	*37*	*23*	39	59	52
18. Personal appearance	03	*32*	*24*	*67*	−08	−03	−10	*40*	−05	−01	11	*69*	−02	18	−06	*23*	*26*	17	51	38	57
7. Observance of rest and lunch periods	04	*29*	*20*	*48*	−13	01	−09	*38*	00	−02	04	*54*	−15	+12	−02	*25*	*25*	16	33	32	36
16. Accepting responsibility for own behavior	05	*38*	*24*	*61*	−10	−17	−09	*50*	07	−02	02	*65*	−03	*26*	04	*22*	*25*	11	43	53	53

Note. The range of residuals was .149 to −.157. Significant factor loadings (those greater than .20) are italicized. Decimals are omitted from all loadings. The items are arranged to maximally disclose simple structure.

[a] Two total scores have the unusually high communalities of 1.00. Since the total scores are a composite of all the variables for each rater (except overall effectiveness), it is not surprising that the factor solution accounts for most of the total score variance.

of stimuli on the same instrument in similar ways.

Factors 4 and 5, accounting for 15% and 7% of the common variance, respectively, demonstrate that raters made many discriminations among the characteristics rated. This indicates that systematic variation exists apart from the rater effect.

Factor 4 has been labeled "competence, skill, and ability." Supervisor ratings of technical competence, ability to organize and schedule work loads, skill in planning nursing care, adaptability in emergencies, dependability, quality of work, completing work on schedule, acceptability of completed work, and attendance and promptness all have meaningful loadings on this factor. Self-ratings load highly on five of these items and peer ratings load highly on six items. The high loadings of supervisor ratings on five additional items indicate that supervisors are less able to distinguish between items reflecting competence and the remaining items than are nurses rated by themselves or their peers.

Factor 5 has been labeled as "effort conformance, and interpersonal qualities." Observance of rules and regulations, effort applied, and impressing visitors have noteworthy loadings on this factor for all raters. In addition, self- and supervisor ratings of personal appearance, observance of rest and lunch periods, and acceptance of responsibility for own behavior load highly on this factor. The factor also includes self- and peer ratings of willingness to perform duties, self-ratings of amount of work performed, and supervisor ratings of communication skills.

DISCUSSION

The results of the factor analysis confirm the evidence cited in the introduction that self-rater, supervisor rater, and peer rater bias exists. This bias is due not only to systematic halo error which may be common to all raters but to rater-group specific bias. The factors which can be interpreted in terms of traits are weak compared to the general and rater-group specific factors. Yet it seems, as Guion (1965, p. 473) suggested, that raters in different positions use different dimensions. For example, supervisors may be less able to discriminate between items related to competence from those related to effort, whereas nurses rating themselves and peers can make this distinction.

Of course, it is possible that the factorial results do not completely reflect the extent of the discriminations made. Since reliabilities of the ratings were not determined, it is impossible to ascertain the uniqueness of each variable. Also, it should be noted that the items on which the subjects were rated and the procedure used to make the ratings were not developed with the intent of maximizing their relevance or reliability.[4] No systematic attempt was made to sample the total performance domain and develop scales that yielded high reliability. This may have resulted in decreasing systematic variance attributable to differences in ratees and increased the systematic error variance attributable to raters.

The factor-analytic approach to multitrait-multirater analysis demonstrated in the present study may help the researcher to explain the factors which are associated with being evaluated differently by different raters. One technique for doing this has been using Ghiselli's (1963) method of examining the off-quadrants in the relationship between the different raters' evaluations (e.g., high rating by supervisor, low rating by peers; low rating by supervisor, high rating by peers; and the same rating given by supervisor and peers) and searching for moderator variables (Lawler, 1967). However, the factor analysis of multitrait-multirater data can aid in this process by uncovering dimensions upon which ratings are made and using these dimensions in the quadrant analysis. For example, high and low supervisor ratings of effort variables may be compared to high and low peer ratings of skill variables on the basis of such moderators as age, tenure, need for achievement, etc. This would result in an even greater understanding of the basis upon which ratings are made.

Since rater bias may be expected, any subjective evaluation of performance should be submitted to dimensional analysis. The hierarchical method allows dimensions which are independent of this bias to be uncovered. Moreover, it may be that different factor scores may be predictive of different criteria. This approach should be helpful in assessment center evaluations or periodic performance appraisals used to predict rate of advancement, future salary increases, or future subjective evaluations for selection or promotion purposes. Similarly where ratings are used as criteria, data from the incumbent supervisor and peers also should be obtained and analyzed. (Subordinate ratings, although not analyzed in this study, may also be useful.) Thus, ratings of the same traits from different points of view may be predictive of different criteria, or themselves be differentially predictive. Which judgments are obtained should depend on their intended use.

---

[4] The authors are grateful to a reviewer for making this point.

REFERENCES

Amir, Y., Kovarsky, Y., & Sharan, S. Peer nominations as a predictor of multistage promotions in a ramified organization. *Journal of Applied Psychology,* 1970, **54**, 462–469.

Barrett, R. S. *Performance rating.* Chicago: Science Research Associates, 1966.

Blum, D., & Naylor, J. *Industrial psychology: Its theoretical and social foundation.* New York: Harper & Row, 1968.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., & Weick, K. *Managerial behavior, performance and effectiveness.* New York: McGraw-Hill, 1970.

Cronbach, L. J. *Essentials of psychological testing.* (3rd ed.) New York: Harper & Row, 1970.

Dunnette, M. D. *Personnel selection and placement.* Belmont, Calif.: Wadsworth, 1966.

Ghiselli, E. E. Moderating effects and differential reliability and validity. *Journal of Applied Psychology,* 1963, **47**, 81–86.

Gordon, L. V., & Medland, F. F. The cross group stability of peer ratings of leadership potential. *Personnel Psychology,* 1965, **18**, 173–177.

Guion, R. *Personnel testing.* New York: McGraw-Hill, 1965.

Harman, H. H., & Jones, W. H. Factor analysis by minimizing residuals (minres). *Psychometrika,* 1966, **31**, 351–368.

Hollander, E. P. Buddy ratings: Military research and industrial implications. *Personnel Psychology,* 1954, **7**, 385–393.

Kavanaugh, M. The content issue in performance appraisal: A review. *Personnel Psychology,* 1971, **24**, 653–668.

Kavanaugh, M., MacKinney, A. C., & Wolens, L. Issues in managerial performance: Multitrait-multimethod analysis of ratings. *Psychological Bulletin,* 1971, **75**, 34–49.

Kirchner, W. K., & Reisberg, D. J. Differences between better and less effective supervisors in appraisals of subordinates. *Personnel Psychology,* 1962, **15**, 295–302.

Lawler, E. E. Management performance as seen from above, below, and within. *Journal of Applied Psychology,* 1967, **51**, 247–253.

London, M., & Klimoski, R. J. Self esteem and job complexity as moderators of attitudes toward work and effectiveness as seen by self, supervisors and peers. Paper presented at the 81st convention of the American Psychological Association, Montreal, 1973. (a)

London, M., & Klimoski, R. J. A study of job complexity: Its effects on performance and satisfaction and a dimensional analysis. Paper presented at the annual meeting of the American Academy of Management, Boston, 1973. (b)

Mayfield, E. C. Management selection: Buddy nominations revisited. *Personnel Psychology,* 1970, **23**, 377–391.

Mayfield, E. C. Value of peer nominations in predicting life insurance sales performance. *Journal of Applied Psychology,* 1972, **56**, 319–323.

Norman, W. T., & Goldberg, L. R. Raters, ratees, and randomness in personality structure. *Journal of Personality and Social Psychology,* 1966, **4**, 681–691.

Oberg, W. Make performance appraisal relevant. *Harvard Business Review,* 1972, **50**, 61–67.

Odiorne, G. S. The need for an economic approach to training. *Journal of American Society of Training Directors,* 1964, **18**(3), 3–12.

Schmidt, F. L., & Kaplan, L. B. Composite versus multiple criteria: A review and a resolution of the controversy. *Personnel Psychology,* 1971, **24**, 419–434.

Schneider, D. E., & Bayroff, A. G. The relationship between rater characteristics and validity of ratings. *Journal of Applied Psychology,* 1953, **31**, 278–280.

Wallace, S. R. Criteria for what? *American Psychologist,* 1965, **20**, 411–417.

Weitz, J. Selecting supervisors with peer ratings. *Personnel Psychology,* 1958, **11**, 25–36.

Wherry, R. J. Hierarchical factor solutions without rotation. *Psychometrika,* 1959, **24**, 45–51.

Wherry, R. J., & Fryer, D. H. Buddy ratings: Popularity contest of leadership criterion? *Personnel Psychology,* 1949, **2**, 147–159.

Whisler, T. F., & Harper, S. F. (Eds.) *Performance appraisal: Research and practice.* New York: Holt, Rinehart & Winston, 1962.

(Received July 5, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 4, 452–455

REALISTIC EXPECTATIONS AS AN AID IN REDUCING VOLUNTARY RESIGNATIONS[1]

DANIEL R. ILGEN[2]
Purdue University

WILLIAM SEELY
United States Army

Realistic information about a situation into which new organizational members were about to enter was presented to the new members after they had decided to join the organization, but before they reported to the organization. When compared to a control group which received no such information, the experimental group showed significantly lower turnover. The results are discussed in reference to the functions served by realistic job information available to new members after they have accepted a position with an organization.

Any organization becomes concerned when apparently committed and carefully selected members voluntarily resign. Typically, attempts are made to improve selection practices and/or to alter the job environment in order to reduce turnover. In a few cases, attempts have been made to alter the beliefs and expectations of prospective organizational members in order to improve the fit between the individuals' beliefs and the organizational environment. Work with insurance agents (Weitz, 1956) and entering cadets at the United States Military Academy (Macedonia, 1969) has demonstrated that applicants, who received realistic information in booklet form about the organization they were considering, more frequently decided to join the organization and less frequently resigned voluntarily. Katzell (1968) also found realistic expectations about the situation reduced turnover among nursing students.

Three explanations seem plausible for the reduction in turnover when applicants receive candid information prior to their decision to join the organization.[3] First, based upon the information in the booklet, some members of the applicant sample may decide not to join the organization. Although some of those who reject the organization at this point may have remained on the job and performed quite well, many who reject the organization may realize that they would eventually find their position unacceptable and resign. Consequently, the sample that joined would contain those individuals who were more likely to remain with the organization.

Second, the applicant's possession of accurate information prior to making a decision to accept or reject the position may lead to a greater felt responsibility for the decision, a greater commitment to the decision, than if the decision were made on the basis of faulty or incomplete information. The increase in commitment to the decision could be reflected in a decrease in the probability of resignation as a possible means of coping with unpleasant aspects of the position. The individual had been forewarned of these unpleasant factors prior to his decision to join the organization.

A third possible reason for the turnover reduction is that knowledge of what to expect from the situation may better prepare individuals to cope with it, and thus reduce frustration. If this is the case, individuals with realistic information about the situation should be better able to cope and less likely to leave the situation than those who have unrealistically positive information about it.

---

[1] The research reported in this article was conducted while both authors were members of the Office of Military Psychology and Leadership at the United States Military Academy, West Point, New York. While the support of the Academy was greatly appreciated, the ideas expressed in this article are solely the opinions of the authors and are not to be considered as the position of the Academy or the U.S. Army.

[2] The authors gratefully acknowledge Robert D. Pritchard for his comments on an earlier draft of this article.

Requests for reprints should be sent to Daniel R. Ilgen, Department of Psychological Sciences, Purdue University, West Lafayette, Indiana 47907.

[3] The insurance agents in Weitz's study received the candid information in booklet form when they applied for the job. Macedonia sent out a booklet to all cadets who had been accepted for admissions two months before the cadets were to report to the Academy. Therefore, Macedonia's sample contained both those who had already decided to join (or not to join) the organization and those who were in the process of making a decision. There was no way to identify these two groups.

The presentation of realistic information prior to the applicant's decision to join the organization confounds the effects of self-selection, increased commitment, and aid in coping. Any or all three of these processes may be operating to reduce turnover. However, information presented *after* the applicant makes the decision to join the organization but before he reports to it should eliminate the first two interpretations. The present study investigated this situation. It was hypothesized that new members would be better prepared to cope with the difficult setting when they know what to expect, even if they were told what to expect after they had decided to enter the organization. Since voluntary resignations were used as an indication of ability to cope with the situation, it was hypothesized that fewer individuals would resign from a group which received this information. Support for this hypothesis would not imply that only aid in coping operates when predecision information is presented, but it would demonstrate that aid in coping is sufficient to reduce turnover in the absence of self-selection and increased commitment to the decision.

METHOD

Setting

The research was conducted with new cadets entering the United States Military Academy in the summer of 1971. New cadets reported on July 1, and spent approximately two months prior to the beginning of the academic year in an intensive training period. This training was designed to familiarize the new cadets with the Academy and with some of the fundamental military disciplines. The period required a quick adjustment on the part of the new cadet and often was very stressful, both psychologically and physically.

Procedure

A booklet was prepared that candidly described the two-month summer-training period and the daily routine during the academic year. This booklet was a revision of a booklet prepared by Macedonia (1969). Material included was based upon interviews and critical incidents solicited from cadets and officers closely associated with the summer program and the day-to-day conditions during the academic year. Care was taken to include mundane and also stressful activities that are often omitted from booklets primarily aimed at attracting individuals to the organization. Drafts of the booklet then were resubmitted to other cadets and officers for ratings of the accuracy and relevance of the material.

The booklet was distributed, by mail during the first week of June, to a random sample equal to one fifth of those individuals who had stated that they would accept their appointments. Of the 246 booklets distributed, one did not reach the cadet due to a change of address, and 11 individuals did not report on the first of July. This left a sample of 234 new cadets who received the booklet. From the remaining sample of new cadets who reported on the first of July but who had not received the booklet, a control group ($n = 234$) was randomly selected.

RESULTS

Voluntary resignations over the two-month summer-training period were recorded for both the experimental group (the group that received the booklet) and the control group. Since some resignations which may have resulted from the summer experience were not processed until a few days after the end of the summer-training session, all resignations before the first day of classes in the fall were considered. During this time period, 27 resigned from the control group and 14 resigned from the experimental group. This difference was statistically significant ($\chi^2 = 4.52$, $df = 1$, $p < .05$) and supported the hypothesis that candid information presented after the decision to participate but before entering the organization reduced the probability of voluntary resignation.

DISCUSSION

Although the data presented supported the hypothesis, the strength of support is tempered by the appropriateness of the control group. The most appropriate control group for this design would be a group equal in size to the experimental group and randomly selected from the applicants who had accepted their appointment by June first. With a control group selected at that time it would have been possible to compare the number of those in each group who had accepted their appointments but did not report on July 1.

However, due to an administrative problem with the study, it was not possible to obtain the most appropriate control group so the reported group was used.

The number of applicants in the experimental group who had accepted appointments but did not report on July 1 could have been less than, the same as, or greater than the corresponding number in the control group. However, only the latter possibility would be damaging to the validity of the interpretation presented. If the number of those who did not report on July 1 were significantly greater for the experimental group than for the control group, the alternative interpretation of self-selection is possible. One could argue that those who did not report after accepting their appointments still had some doubts, and the booklet convinced them not to come. The "doubters" were still in the control on July 1 but subsequently resigned, thus creating the significant difference reported.

A consideration of the Macedonia (1969) data and one additional analysis led the authors to conclude that the possibility of a significantly smaller number of the applicants who did not report in the control group than in the experimental group was highly unlikely. Macedonia (1969) found that a significantly *greater* number of cadets reported the first of July from the group that received a similar booklet than from the group that did not receive the booklet. His sample included both those who had and those who had not accepted their appointments. In his sample, it was not possible to separate those who had accepted their appointments from those who had not. However, it is highly unlikely that the significant difference was due primarily to those who had accepted their appointments. In a similar vein but with a less comparable sample, Weitz (1956) found that it was easier to fill insurance agent positions when candid information was presented to applicants than when it was not.

An additional set of analyses applied successive chi-squares to discover the critical number of persons necessary in order to have the number not reporting in the experimental group significantly greater than that of the control. Of the 245 who were sent booklets, 11 did not report July 1. Assuming an equal-sized control group had been used ($n = 245$), in order to obtain a significant difference at the $p < .05$ level between the number not reporting to the control group and the 11 not reporting in the experimental group, only three or less of those in the control group would have had not to report. Only three of 245 over a one-month period seemed highly unlikely in light of the normal attrition prior to the July 1 reporting date and in light of the Macedonia data. Therefore, the authors were willing to accept the interpretation presented as the most plausible one.

Finally, the question of the degree to which these results may generalize to other settings should be addressed. Most new cadets prior to their first day at the Academy felt that the summer training was going to be very difficult, although they had many misconceptions about specific aspects of the training. The information in the booklet spelled out specific factors of which they may not have been aware, but this information did not contradict their previous general impression of the training. We would hypothesize that the generalizability of this study may be influenced by the degree to which candid information about the job fits the general impressions of the newly accepted applicants. Perhaps, if there is a large discrepancy between the information provided after the decision to join the organization and what the applicant was led to believe when he accepted the position, the new employee may feel he has been deceived and candid information may not reduce voluntary resignations.

CONCLUSION

The effect of realistic expectations, established after the decision to join the organization had been made, demonstrated that reduced resignations need not be due to increased commitment to the decision or to self-selection by the applicants. Self-selection has been the concern of those whose primary interest is recruiting. Recruiters believe that potentially valuable organization members who would remain with the organization may be turned away by the initial information if that information does not present a very positive picture. Although both Weitz (1956) and Macedonia (1969) showed that it was easier to attract new members if the new members received candid information, and Macedonia found no significant differences on variables considered for acceptance (e.g., high school rank, Scholastic Aptitude Test scores, etc.) between those who declined their appointments and those who accepted their appointments. Organizations often are reluctant to explicitly describe positions to applicants. However, the data from the present study indicated that even if one is reluctant to describe the position completely prior to the applicant's decision to join the organization, new members can benefit from this information if they receive it *after* their decision to join. This suggests efforts to aid new members in their adjustment to the job should begin even before they report to work.

REFERENCES

Katzell, M. E. Expectations and dropouts in schools of nursing. *Journal of Applied Psychology*, 1968, **52**, 154–158.

Macedonia, R. M. Expectation, press, and survival. Unpublished doctoral dissertation, New York University, 1969.

Weitz, J. Job expectancy and survival. *Journal of Applied Psychology*, 1956, **40**, 245–247.

(Received July 16, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 4, 456–464

PROBABILITY AND UTILITY ASSUMPTIONS UNDERLYING USE OF THE STRONG VOCATIONAL INTEREST BLANK

FRANK L. SCHMIDT[1]

Michigan State University

Decision implications of Strong's assumption of equal base rates for his men-in-general (MIG) and individual occupational groups are explored. Using the Strong Vocational Interest Blank (SVIB) Physician's scale as an example, it is demonstrated that when realistic base rates are assumed, total decision error rate using the SVIB—even when optimal cutting scores are employed—is at best about equal to that resulting from prediction using base rates alone. It is shown that use of the SVIB as recommended by Strong implicitly assumes that false negative errors are much more serious than are false positives. Finally, it is shown that when "reasonable" utility or gain matrices are assumed, use of the SVIB can reduce total error rate despite the limitations imposed by low occupational base rates, and that this effect is greater when statistically optimal cutting scores, rather than those recommended by Strong, are used.

In many practical decision-making tasks faced by psychologists, scores on some continuous test or variable are used to make dichotomous decisions. In a study that is now a classic, Meehl and Rosen (1955) demonstrated that cutting scores used in such decisions must take into account population base rates, that is, the relative frequency of the category or group of interest. They showed that a device which usefully reduces decision or classification errors when the base rate of the dichotomous phenomenon of interest (e.g., schizophrenic versus normal or shoplifter versus honest shopper) is near .50 may be useless—or may even lead to increased errors—when the base rate is very high or very low. They presented conditions that must be met for the test to achieve a greater degree of accuracy than would result from use of the base rates alone. Dawes (1962) later provided a more easily understood analysis in terms of conditional probabilities. Rorer, Hoffman, LaForge, and Hsieh (1966) presented an analytic (as opposed to empirical) solution to the determination of optimal cutting scores, and Rorer, Hoffman, and Hsieh (1964) provided extensive tables of optimum cutting scores for groups differing both in base rate and variance. These tables assume that all errors are equally costly and all correct classifications are equally beneficial; a subsequent study (Rorer, Hoffman, & Hsieh, 1966) presented a method of determining optimal cutting scores when false positives and negatives are unequal in cost and true positives and negatives are unequal in benefit.

The decision theory principles explored in these studies can be applied to an evaluation of the Strong Vocational Interest Blank (SVIB)—perhaps the most commonly used instrument in occupational counseling. The purpose of the present study is to show that application of these principles leads to conclusions about the usefulness of the SVIB that are quite different from those reached by Strong (1955), Strong and Campbell (1966), and Campbell (1968, 1971). The following analysis applies equally to the 1966 revision and to the earlier form. Similarly, while it is in terms of the men's form, the analysis generalizes easily in all its essentials to the women's form of the SVIB.

IMPACT OF BASE RATES ON EXPECTANCIES AND ON TOTAL ERROR RATES

Meehl and Rosen (1955) and Dawes (1962) have pointed out that criterion validation all too often results in a statement of test validity in terms of $P(R_j/G_i)$, that is the probability of Response j for subjects in Group i (e.g., the probability that schizo-

---

[1] Requests for reprints should be sent to Frank L. Schmidt, Department of Psychology, Michigan State University, East Lansing, Michigan 48824.

The author is grateful to John Hunter and Robert Forsythe for helpful comments on an earlier draft of this article.

phrenics will score above 70 on a test). The desired index of predictive efficiency, however, is $P(G_i/R_j)$, the probability of group membership, given the response (e.g., the probability that the respondent is a schizophrenic given that he scores above 70). These two probabilities are related by means of Bayes' theorem:

$$P(G_i/R_j) = \frac{P(R_j/G_i)P(G_i)}{P(R_j)}.$$

Strong (1955) successfully avoided this common error and presented validity evidence for the SVIB in the form $P(G_i/R_j)$, where G_i is the occupation in question and R_j is a letter grade on the SVIB scale for the occupation in question. In summarizing the results of his 18-year follow-up study of 884 Stanford University students, Strong (1955, p. 43) presented the probabilities (expectancies) shown in Figure 1 as averages across 16 occupations. The probability of being in a given occupation given that one obtained an A−, for example, on the scale for that occupation 18 years earlier is .74. These probabilities are quite impressive, and this summary is widely quoted as evidence of the SVIB's effectiveness (e.g., Anastasi, 1968, p. 472; Super & Crites, 1962, p. 440). But, as pointed out by Dolliver (1969), such evidence is both erroneous and highly misleading.

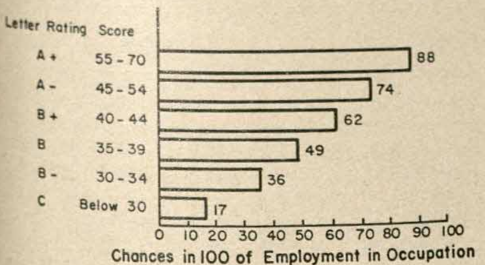


FIGURE 1. Chances in 100 of employment in occupation given specific letter grade on the Strong Vocational Interest Blank 18 years earlier (Strong, 1955, p. 43).

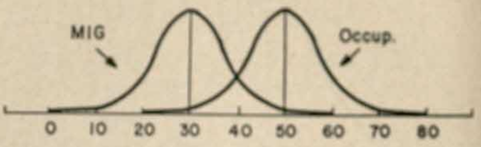


FIGURE 2. Distributions of the men-in-general and specific occupation groups under the assumption of equal base rates.

Each SVIB scale is designed to discriminate maximally between that occupation and the men-in-general (MIG) group, a large heterogenous group of men in various professional and paraprofessional occupations. Because base rates in Strong's (1955, pp. 40–41) MIG group were different for different occupations and because Strong felt that within occupations, these base rates would vary as between universities and geographical regions, he decided to express expectancy ratios for all scales "in terms of samples of equal size." The implications of this procedure can be seen by comparing Figures 2 and 3.[2] In comparing each occupation with the MIG group, Strong assumed the situation shown in Figure 2, that is, an occupational group equal in size to the MIG group. This assumption is obviously highly unrealistic; none of the occupations scaled on the SVIB even approximates the relative frequency of all other occupations combined—and this holds true whether we consider base rates in Strong's MIG group or in the United States male employed professional-technical population. Figure 3 illustrates a more reasonable assumption; it depicts the base rate for the occupation as being much smaller than that of the nonoccupation, or MIG group. Operating on the basis of this assumption leads one to estimates of $P(G_i/R_j)$ much smaller than those shown in Figure 1, with the degree of reduction in size depending on the estimate chosen of the base rate of the occupation. As an example, consider the physician's scale. Physicians constituted 6% of the original MIG group (Strong, 1955, p. 68); they make up 5.1% of the United States employed male professional population (U.S. Bureau of the Census, 1963).[3] When considered in relation

---

[2] Scaling of the abscissa in Figures 2 and 3 approximates that of the typical SVIB scale (i.e., $\bar{X}_{\text{MIG}} \cong 30$, $\bar{X}_{\text{occup.}} \cong 50$, and $SD_{\text{MIG}} \cong SD_{\text{occup.}} \cong 10$).

[3] 1960 census figures; at the time of this writing, national 1970 census breakdowns by occupational group were not yet available. But it is unlikely that the 1970 base rates will be significantly different from 1960 figures.

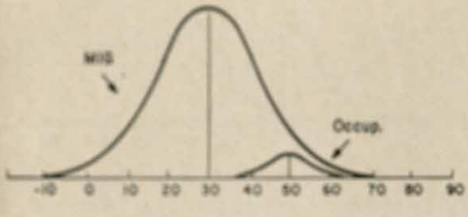


FIGURE 3. Distributions of the men-in-general and specific occupation groups under the assumption of unequal base rates.

to the combined U.S. employed male professional, managerial, sales, and farm populations—all of which occupational categories are included in the SVIB MIG group—the physician's base rate is only 1.6% (U.S. Bureau of the Census, 1963). In certain population subgroups, of course, base rates may be higher. For example, in the population of undergraduate biology majors the proportion of future physicians may be as high as .20.

Assuming a 20-point mean difference between physicians and the MIG group (Strong, 1955, p. 37) and the standard deviations of 13 and 10 for the MIG and physician's groups, respectively (Strong, 1955, p. 68), application of Bayes' theorem[4] yields the expectancies shown in Table 1 for the base rates of .50 and .20, and the rounded base rates of .05 and .02. (The base rate .060 was excluded because of its closeness to .051.)

The physician's scale discriminates somewhat better than the average of all SVIB scales, and the result is a very impressive set of expectancies under the assumption of equal base rates. For example, a college student scoring A− is given 8 chances out of 10 of being a physician 18 years later. However, these expectancies drop off sharply when more realistic base rates are entered into the equation. When the base rate is taken as .20, a student scoring A− has even chances of not becoming a physician. When .05, rather than the .02 which is perhaps more realistic in most situations, is employed as the base rate, none of the expectancies is as large as .50. No matter how high the student scores, the probability is always much higher that he will become a nonphysician than a physician. This problem is not peculiar to the occupation of physician. In fact, most of the occupations on the SVIB that could be found in the census figures showed smaller base rates than the physician category.

Another approach to this problem is via an examination of the total error rate—the percent of all decisions that are in error—as

TABLE 1

$P(G_i/R_j)$ FOR THE STRONG VOCATIONAL INTEREST BLANK PHYSICIAN'S SCALE UNDER FOUR BASE RATE ASSUMPTIONS

Letter rating	Score	Base rate			
		.50	.20	05	.02
A+	55–70	.92	.74	7	.19
A−	45–54	.80	.50	6	.07
B+	40–44	.60	.28	2	.03
B	35–39	.41	.15	4	.02
B−	30–34	.21	.07	[illegible]	.01
C	Below 30	.04	.01	2	.001

Note. Essentially, $P(G_i/R_j)$ is the probability of a man's being a physician, given that he obtained a certain score on the physician's scale 18 years earlier.

TABLE 2

PERCENTAGE OF ALL DECISIONS INCORRECT FOR STRONG VOCATIONAL INTEREST BLANK PHYSICIAN'S SCALE AS FUNCTION CUTTING SCORE AND BASE RATE

Cutting score	Base rate			
	.50	.20	.05	.02[d]
45 (A−)	21.7	16.2	13.4	12.9
40 (B+)	19.0	20.8	21.2	21.9
40.13[a]	18.5	—	—	—
49.5[b]	—	15.7	—	—
65.39[c]	—	—	4.9	—

[a] Optimal cutting score for base rate = .50.
[b] Optimal cutting score for base rate = .20.
[c] Optimal cutting score for base rate = .05.
[d] For base rate = .02, there is no cutting score that results in fewer errors than prediction using base rate alone.

---

[4] Values inserted into the right side of Bayes' equation were calculated as follows: $P(R_j/G_i)$ was determined for the physician's distribution for each letter grade using normal curve tables. $P(G_i)$ is, of course, the assumed physician's base rate. $P(R_j) = P(R_j/G_i) \cdot P(G_i)$ plus the probability of the response in the MIG group times the assumed base rate of the MIG group, that is, $1 - P(G_i)$. These calculations, and those for Table 2, assume normality of both MIG and Occupation scale distributions. Although there are some apparent departures from normality in some occupational scales, they are not large enough to significantly affect the figures in Tables 1 and 2.

a function of cutting score and assumed base rate. Strong (1955, p. 54) has advocated that students "seriously consider" an occupation on which they score B+ or above, thus advising counselees to tentatively classify themselves into such an occupation rather than into the MIG group. The more stringent cutting score of A− is also sometimes used. In addition, the statistically optimal cutting score can be determined for each assumed base rate. Table 2 presents total error rates for these cutting scores as a function of four base rates.[5] When a base rate of .50 is assumed, Strong's recommended cutting score of B+ is close to optimal; in addition, all cutting scores reduce the error rate considerably below that resulting from use of the base rate alone (50%). When the base rate is taken as .20, A− is close to optimal as a cutting score, while the B+ suggested by Strong produces an error rate above the 20% base-rate level. When the base rate is taken as .05, both letter grade cutting scores produce error rates considerably above the 5% level that would result from use of the base rate alone. Even the optimal cutting score reduces base rate error by only .1%. When .02 is taken as the base rate, no cutting score can be drawn which produces fewer errors than does the base-rate prediction that no one will become a physician (2% error rate). Even if we are willing to assume a .05 base rate for future physicians in most counselee populations, the reduction of .1% in total error hardly makes use of the SVIB appear worthwhile. Clearly, if use of the Strong can be justified, justification must be on the basis of inequality (i.e., differential weighting) of classification errors.

INEQUALITY OF ERRORS

The SVIB is probably most realistically viewed as one step in a sequential career decision-making process. The individual counselee will typically obtain C or B− ratings on most of the 54 occupational scales, and most of these occupations are usually then tentatively eliminated from further consideration. The occupation that is in fact optimal for the counselee will probably be found among the relatively small number of scales on which he scores A or B+, and these occupations are thus retained for closer scrutiny. If we assume that the individual can then eliminate the inappropriate occupations (scales) in this group by some other means (e.g., financial or practical considerations, personal insight or preference, consideration of intellectual ability requirements, etc.), then false positives become less important than false negatives. The cost of false negatives is higher because, once assigned to the MIG group on a given scale, the counselee does not ordinarily investigate that occupation further, thus making the error a potentially permanent one. False positive errors, on the other hand, are usually remedied when the individual more closely scrutinizes the occupations on which he obtains an A or B+ rating.

Given that the two different types of error are unequal in cost, the next question is whether correct classifications, true positives and negatives, are equal or unequal in value. Practical considerations, for example, the fact that for each individual there are many true negatives but only one true positive; the fact that many true negatives can be correctly arrived at by the individual without benefit of the SVIB, etc., seem to indicate that true positives are of greater value than true negatives. Nevertheless, it might be argued that there are three sets of assumptions about true positives and negatives that are plausible. Assumption 1 states that gain in utility from a true positive (G_{++}) and gain from a true negative (G_{--}) are both zero; true positives and negatives merely function (and function equally) to avoid losses in utility that result from false positives and negatives. Assumption 2 holds that true positives and negatives produce equal positive gains. Assumption 3 assumes that both gains in utility are positive, but that the value of a true positive exceeds that of a true negative. These three assumptions can be expressed algebraically as follows:

---

[5] For the physician's scale, $\bar{X} = 50$, $SD = 10$; for the MIG distribution, $\bar{X} = 30$, $SD = 13$ (Strong, 1955, p. 68). For each cutting score, total error rate is percent of false positives in the MIG group (determined from normal curve tables) times the assumed MIG base rate plus the percent of false negatives in the physician's group (also from normal curve tables) times the assumed physician base rate. Optimal cutting scores for all base rates except .20 were calculated by interpolation using the Rorer, Hoffman, and Hsieh (1964) tables. Since the base rate of .20 is not contained in these tables, this optimal cutting score was calculated using Equation 1 of Rorer, Hoffman, LaForge, and Hsieh (1966).

1. $G_{++} = G_{--} = 0$
2. $G_{++} = G_{--} > 0$
3. $G_{++} > G_{--} > 0$.

Sets of assumptions other than those three (e.g., $G_{++} < G_{--}$ and $G_{++} = G_{--} < 0$) can be eliminated as implausible. The decision principles presented by Rorer, Hoffman, and Hsieh (1966) can now be used to shed light on the *extent* of inequality that is implicitly being assumed between false positive and false negative errors under each of these three sets of assumptions when the SVIB is used in counseling as recommended by Strong (1955), Strong and Campbell (1966), and Campbell (1968, 1971). That is, we can recover some of the properties of the gain or utility matrix implicitly assumed by Strong (1955).

Given a gain or utility matrix such as that shown in Table 3, Rorer, Hoffman, and Hsieh (1966) showed that

$$\frac{B}{1-B}\cdot\frac{G_{++}-G_{+-}}{G_{--}-G_{-+}} = \frac{B'}{1-B'}, \qquad [1]$$

where B = the true base rate for the occupational group and B′ = the base rate adjusted for the inequalities in the gain or utility matrix. B′ is the base rate needed to enter the table of optimal cutting scores, given a particular B and utility or gain matrix. For convenience let

$$\frac{G_{++}-G_{+-}}{G_{--}-G_{-+}} = G.$$

TABLE 3

GAINS MATRIX FOR AN INDIVIDUAL OCCUPATIONAL SCALE ON THE STRONG VOCATIONAL INTEREST BLANK

Actual	Predicted	
	Men-in-general (−)	Occupation (+)
Men-in-general (−)	G_{--}	G_{-+}
Occupation (+)	G_{+-}	G_{++}

Note. G = gain.

Then, solving for G in Equation 1 gives

$$G = \frac{B'(1-B)}{(1-B')B} \qquad [2]$$

In computing all expectancy ratios on the basis of equal base rates, Strong was in effect assuming B′ = .50. Using this value for B′, Equation 2 yields G = 4 when B = .20, G = 19 when B = .05, and G = 49 when B is taken as .02. Then under Assumption 1 that $G_{--} = G_{++} = 0$,

$$G_{.20} = \frac{G_{+-}}{G_{-+}} = \frac{4}{1}$$

$$G_{.05} = \frac{G_{+-}}{G_{-+}} = \frac{19}{1}$$

and

$$G_{.02} = \frac{G_{+-}}{G_{-+}} = \frac{49}{1}$$

where .20, .05, and .02 refer to base rates. Thus, if true positives and negatives are considered to each represent zero gain, then use of the SVIB physician's scale assumes that a false negative is 4, 19, and 49 times as harmful as a false positive error if the true base rate is taken as .20, .05, and .02, respectively. (The reader should recall that both G_{+-} and G_{-+} are negative numbers.)

If we assume $G_{--} = G_{++} > 0$, then when the base rate is .20,

$$G_{.02} = \frac{G_{++}-G_{+-}}{G_{--}-G_{-+}} = 4.$$

Substituting G_{++} for G_{--} and solving, we get

$$\frac{G_{+-}}{G_{-+}} = 4 - \frac{3G_{++}}{G_{-+}}.$$

Since G_{-+} is a negative number (a negative "gain"), as G_{++} and G_{--} increase a false negative comes to be considered *more* than four times worse than a false positive. Similarly, when the base rate is taken as .05,

$$\frac{G_{+-}}{G_{-+}} = 19 - \frac{18G_{++}}{G_{-+}}.$$

And, finally, when the base rate is .02, we get

$$\frac{G_{+-}}{G_{-+}} = 49 - \frac{48G_{++}}{G_{-+}}.$$

As G_{++} and G_{--} increase, false negatives become more costly relative to false positives. Thus, if we consider true positives and negatives to have positive and equal value, use of the SVIB physician's scale as advocated by Strong (1955) assumes that false negatives are more than four times as serious as false positives when the base rate is taken as .20. When the base rate is taken as .05, use of the SVIB assumes that false negatives are more than 19 times worse than false positives. When the base rate of .02 is used—the most realistic of the three base rates for most high school and college counselee populations—the implicit assumption is that classifying a counselee into the MIG group when he will in fact enter the occupation is more than 49 times as serious as the opposite error.

Finally, if we assume (Assumption 3) that $G_{++} > G_{--} > 0$, then, when the base rate is taken as .20,

$$G_{.20} = \frac{G_{++} - G_{+-}}{G_{--} - G_{-+}} = 4.$$

Solving for G_{+-}/G_{-+}, we get

$$\frac{G_{+-}}{G_{-+}} = 4 + \frac{G_{++} - 4G_{--}}{G_{-+}}.$$

Since G_{-+} is a negative number, G_{+-}/G_{-+} will be less than 4 only when $G_{++} > 4\ G_{--}$. For all other permissible values of G_{++}, a false negative will be considered more than four times as serious as a false positive. When the base rate is taken as .05,

$$\frac{G_{+-}}{G_{-+}} = 19 + \frac{G_{++} - 19G_{--}}{G_{-+}}.$$

Here the value of a true positive (G_{++}) must be considered to be more than 19 times the value of a true negative (G_{--}) for the cost of a false negative (G_{+-}) to be considered less than 19 times as great as a false positive (G_{-+}). When the base rate is .02,

$$\frac{G_{+-}}{G_{-+}} = 49 + \frac{G_{++} - 49G_{--}}{G_{-+}}.$$

With this, the most realistic of the base rates, the cost of a false negative (G_{+-}) is considered greater than 49 times the cost of a false positive (G_{-+}) unless the value of a true positive (G_{++}) is considered to be equal to or greater than 49 times the value of a true negative (G_{--}).

On logical and rational grounds, both Assumption 3 and a base rate of .02 appear the most realistic for most situations in which the SVIB is used. On the same grounds, it would appear unlikely that true positives are in fact considered more than 49 times as useful as true negatives. The conclusion indicated, then, is that use of the SVIB physician's scale as recommended by Strong (1955, p. 54) implicitly assumes that false negative errors are approximately 49 times as harmful as false positive errors.[6]

ACCURACY GIVEN A "REASONABLE" GAINS MATRIX

Strong's (1955) assumption that the adjusted base rate (B′) is .50 leads backward to implicit gains matrices with G = 4, 19, and 49 when the true base rate (B) is taken as .20, .05, and .02, respectively. Ordinarily, of course, the process is reversed: The four values in the gains matrix are specified, and these values, along with B, are entered into Equation 1 to solve for *B′*. *B′* is then used to enter the Rorer, Hoffman, and Hsieh (1964) tables to find the optimum cutting score and the adjusted total error rate. An important question then, is whether, given gains matrices containing "reasonable" values, use of the SVIB leads to an adjusted total error rate lower than *B′*, the error rate obtaining when prediction is on the basis of B′ alone. The

---

[6] Some counselees may expect the SVIB to indicate not the specific occupation that would be optical for them but rather a cluster of family of occupations, most of which would in practice be suitable from the viewpoint of interests. When the SVIB is used this way, the appropriate unit of analysis would be the 11 occupational families on the SVIB. With this kind of analysis, the base-rate problem would be reduced (few families would have base rates as low as .02), and the number of decision errors, especially false positives, would be reduced. While this may be a realistic model for the use of the SVIB for many counselees, especially in view of the high occupational scale intercorrelations within most occupational families, it does not seem to be that of Strong (1955). Strong's focus is on the individual occupations. Nevertheless, a subsequent study is planned using the occupational families as the unit of analysis.

TABLE 4

A "REASONABLE" GAINS MATRIX FOR INDIVIDUAL STRONG VOCATIONAL INTEREST BLANK SCALES

Actual	Predicted	
	Men-in-general (−)	Occupation (+)
Men-in-general (−)	+1	−1
Occupation (+)	−10	+10

gains matrix shown in Table 4 might be reasonable for many counselees. Table 4 assumes that correct identification of the individual's future occupation (Cell ++) is 10 times as useful as the correct identification of the occupation as inappropriate for the individual (Cell −−). Similarly, incorrect rejection of the occupation as inappropriate (Cell +−) is 10 times as harmful as incorrectly labeling the occupation as appropriate (Cell −+).[7] In an individual decision-making process of this sort, actual gains matrices would, of course, vary from counselee to counselee, depending on individual differences in value systems (Cronbach & Gleser, 1965, pp. 7–9). For other counselees, a "reasonable" gains matrix might be produced by substituting +20 for the +10 in cell G_{++} of Table 4 and a −20 for the −10 in cell G_{+-}. Such counselees would consider false negatives more costly and true positives of greater value than individuals characterized by the gains matrix in Table 4. This we refer to as the second or alternative "reasonable" gains matrix.

For both these "reasonable" gains matrices, Table 5 shows for each of three base rates: (*a*) B′ (from Equation 2), (*b*) total adjusted error rate,[8] and (*c*) percent reduction in error using (*a*) optimal cutting scores (from Equation 1 of Rorer, Hoffman, & Hsieh, 1966) and (*b*) Strong's (1955, p. [illegible]) suggested cutting score of 40 (B+). When the base rate is taken as .20, use of the SVIB reduces total adjusted error rate below levels that would result from prediction using B′ only, but these gains are considerably larger when optimum cutting scores, rather than Strong's suggested score, are employed. This difference is especially large for the alternative gains matrix. In this instance the reduction in total error that would result using Strong's cutting score is so small (.5%) as to be negligible.

When the base rate is taken as .05, the cutting score advocated by Strong leads to a significant reduction in total error for the first gains matrix and is only marginally less efficient than the optimal cutting score. For those counselees whose values correspond to the alternative gains matrix, Strong's cutting score is essentially identical to the optimal cutting score at this base rate. Under this set of circumstances, procedures recommended by Strong (1955, p. 54) for the use of the SVIB are ideal.

The optimal cutting score, when B = .02, works out to 51 for the first gains matrix and produces a total adjusted error rate of 13.2% (versus the 17% expected when predicting using adjusted base rate alone). This constitutes a 22.4% reduction in error. However, when the cutting score of 40 recommended by Strong is used instead of the optimum cutting score of 51, total adjusted error rate is 21.0%—a greater error rate than that which obtains when predictions are made

---

[7] These statements, and the analyses of the previous section, assume measurement of the utilities on a ratio scale. However, as pointed out later, empirical estimation of the utility of the SVIB for a given individual requires measurement of the four utilities in his gains matrix only on an interval scale. As can be seen in Equation 1, this follows from the fact that it is the ratio of differences rather than the direct ratio of utilities that is used in computing B′.

[8] Both total adjusted error and percent reduction in error are, of course, averages that would be expected over a large number of individuals with identical gains matrices. The model used in this study—which is general statistical decision theory with differential utilities—was developed for situations in which many decisions are made under similar circumstances. In the case of individual decisions, each individual's gains matrix is potentially unique (though it is likely in practice that some individuals will have essentially identical matrices). When applied as above, the model implicitly assumes, for each gains matrix, a hypothetical population of individuals all with the same gains matrix; it then gives the average error that would obtain over this population. For any individual with this gains matrix, this is the best estimate of the value of the SVIB for him. This estimate must be based on a model which assumes a number of similar decisions; each individual's decision is a unique event which never recurs and thus cannot be modeled in this sense.

TABLE 5

B′, TOTAL ADJUSTED ERROR RATE, AND PERCENT REDUCTION IN ERROR FOR TWO "REASONABLE" GAINS MATRICES AT THREE BASE RATES (B)

B	B′	Optimum cutting score			Strong's (1955) cutting score		
		Cutting score	Total adjusted error rate	% reduction in error	Cutting score	Total adjusted error rate	% reduction in error
"Reasonable" gains matrix in Table 4 (G = 10)							
.20	.71[a]	34.4	15.0	48.3	40	17.7	39.0
.05	.34	44.0	18.6	45.3	40	20.2	40.5
.02	.17	51.0	13.2	22.4	40	21.0	−23.5
Alternative "reasonable" gains matrix (G = 20)							
.20	.83[a]	30.8	10.4	38.8	40	16.9	.5
.05	.51[a]	40.4	18.9	61.4	40	18.9	61.4
.02	.29	45.7	17.7	39.0	40	20.3	30.0

[a] Since $B' > .50$, error rate for base-rate prediction is $1 - B'$.

from the adjusted base rate alone. On the other hand, when the alternative gains matrix is assumed, use of Strong's cutting score leads to a useful reduction in total decision error, though not as great a reduction as is produced by the optimal cutting score. Of the three base rates examined in the present context, .02 is probably the most realistic for most of the high school and college counselee populations with which the SVIB is used. Utility of this instrument when used with Strong's suggested cutting score in these populations will be negative for those individuals whose subjective gains matrix corresponds to the matrix in Table 4. As the counselee's gains matrix moves away from the Table 4 matrix and toward our alternative gains matrix, the positive utility of the SVIB for him increases.

In conclusion, when the utility of the SVIB is calculated using more realistic base rates, the instrument still shows a positive utility in many situations, given a "reasonable" gains matrix. Utilities are highest, however, when optimal cutting scores rather than those recommended by Strong are used, and under certain, possibly common, combinations of base rate and gains matrix, use of Strong's cutting score can lead to greater error rates than are produced by prediction from base rates alone.

CONCLUSION

In summary, the SVIB, if carefully used, probably has a significant positive utility for most individuals, despite the limitations on predictive efficiency imposed by low occupational base rates. This utility, however, can be cancelled or even rendered negative under some circumstances if nonoptimal cutting scores—including those recommended by Strong—are employed. Whether or not this latter effect occurs depends directly on the relative negative values placed on false positives and negatives and the relative positive values placed on true positives and negatives, that is, on the gains matrix of the individual. These conclusions point to a need for the measurement of actual counselee gains matrices.[9] If gains matrices of individual counselees could be assessed, optimal cutting scores and expected utility of the SVIB could be assessed for each individual.

---

[9] Measurement problems in the present context are not as complex as they might seem at first glance. The four utilities in an individual's gains matrix need to be measured on only an interval scale. A ratio scale, that is, a scale with a rational zero point, is not necessary (see footnote 6). Standard psychometric procedures are generally considered to give adequate approximations to interval scales (Nunnally, 1967, chap. 1).

REFERENCES

Anastasi, A. *Psychological testing.* New York: Macmillan, 1968.

Campbell, D. P. The Strong Vocational Interest Blank: 1927–1967. In P. McReynolds (Ed.), *Advances in psychological assessment.* Vol. 1. Palo Alto, Calif.: Science & Behavior Books, 1968.

Campbell, D. P. *Handbook for the SVIB.* Stanford, Calif.: Stanford University Press, 1971.

Cronbach, L. J., & Gleser, G. C. *Psychological tests and personnel decisions.* Urbana: University of Illinois Press, 1965.

Dawes, R. M. A note on base rates and psychometric efficiency. *Journal of Consulting Psychology,* 1962, **26**, 422–424.

Dolliver, R. H. "3.5 to 1" on the Strong Vocational Interest Blank as a pseudo-event. *Journal of Counseling Psychology,* 1969, **16**, 172–174.

Meehl, P. E., & Rosen, A. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin,* 1955, **52**, 194–216.

Nunnally, J. C. *Psychometric theory.* New York: McGraw-Hill, 1967.

Rorer, L. G., Hoffman, P. J., & Hsieh, Kuo-Cheng. Tables of optimum cutting scores to discriminate groups of unequal size and variance. Eugene, Oregon: Oregon Research Institute Technical Report, 1964, **4**, (3).

Rorer, L. G., Hoffman, P. J., & Hsieh, Kuo-Cheng. Utilities as base rate multipliers in the determination of optimum cutting scores for the discrimination of groups of unequal size and variance. *Journal of Applied Psychology,* 1966, **50**, 364–368.

Rorer, L. G., Hoffman, P. J., La Forge, G. E., & Hsieh, Kuo-Cheng. Optimum cutting scores to discriminate groups of unequal size and variance. *Journal of Applied Psychology,* 1966, **50**, 153–164.

Strong, E. K. *Vocational interests 18 years after college.* Minneapolis: University of Minnesota Press, 1955.

Strong, E. K., Jr., & Campbell, D. P. *Manual for Strong Vocational Interest Blank.* Stanford, Calif.: Stanford University Press, 1966.

Super, D. E., & Crites, J. O. *Appraising vocational fitness.* New York: Harper & Row, 1962.

U.S. Bureau of the Census. *U.S. census of the population: 1960. Detailed characteristics. United States Summary.* (Final Report PC(1)-1D) Washington, D.C.: U.S. Government Printing Office, 1963.

(Received June 7, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 4, 465–471

CONSUMER DISSATISFACTION AS A FUNCTION OF DISSATISFACTION WITH INTERPERSONAL RELATIONSHIPS

GERHARD W. H. SCHERF[1]

University of Guelph, Ontario, Canada

An application of Alderfer's existence, relatedness, and growth (E.R.G.) theory investigated whether frustrated relatedness needs resulted in additional consumer existence needs which, in turn, produced consumer dissatisfaction. Studies of 306 randomly chosen subjects indicated that a decrease of consumers' ratings of satisfaction in their relations with "significant others" and with people at the place of employment was correlated with an increasing desire for more items of material value in the consumer role. Some correlations reached higher statistical significance when respondents of the lowest income group and those of both the oldest and the youngest age categories were eliminated from the sample. Higher incomes correlated positively with existence need satisfaction. The results supported Alderfer's theory.

Many cases of consumer dissatisfaction can be traced to causes in the marketplace, such as instances of malfunctioning of products or of discoveries of illegitimate pricing. However, much of today's discontent with matters of consumption may possibly be viewed as resulting from frustration in life roles other than the consumer role. Support for this proposition comes from Hulin (1969), whose studies of job satisfaction and its relatedness to general living conditions in the community have drawn attention to the possibility that satisfaction and/or dissatisfaction in one life role may be transferred into other life roles. The application of this principle to the current problem of consumer dissatisfaction, however, is notably absent in the literature. Yet, in view of what is known about the various phenomena of adjustment to frustrations, particularly projection and scapegoating, it is quite possible that consumer satisfaction and/or dissatisfaction originates from satisfaction or frustration experienced in various life roles, not merely in the consumer role. In fact, personnel working in consumer complaint centers could perhaps testify to the face validity of this proposition.

On a theoretical basis, such a cause/effect relationship for consumer dissatisfaction could also be constructed from the concepts of Alderfer's (1969) E.R.G. theory; this conceptual framework classifies human needs into three categories: (*a*) existence, (*b*) relatedness, and (*c*) growth.

Existence needs include all goods and services offered in the marketplace. In accordance with the E.R.G. theory, existence needs, the lowest of the need categories, are ordinarily pursued only to the point where the individual's physiological and psychological survival is assured. In our North American setting, this state can perhaps be reached by a majority of consumers with relatively little effort. Simultaneously with the pursuance of existence needs and predominantly after the existence needs are met, the individual works at the satisfaction of the next higher level of needs: relatedness needs.

Alderfer defines relatedness needs as needs involving relationships with significant other people. People are assumed to satisfy relatedness needs by mutually sharing their thoughts and feelings. If the social environment is favorable, the relatedness needs will soon be relatively satisfied, and the individual can devote his time and effort to the pursuance of growth needs. However, if the social climate is unfavorable, the pursuance of relatedness needs may soon lead to a state of frustration. As a result of his being unable to satisfy his relatedness needs, the individual may then regress to the creation and pursuance of additional existence needs in the form of expanded consumer wants. (The E.R.G. theory's Proposition 2: The more relatedness needs are frustrated, the more existence needs will be desired.) Since most consumers' income is inevitably limited, dissatisfaction is likely to arise soon out of their inability to satisfy all of the newly created existence needs.

---

[1] The author wishes to express his sincere appreciation to Wilhelmina Owens who collected data and to Susan Kimpel who contributed to all aspects of the project in her capacity as research assistant. Thanks are also extended to the members of the Research Advisory Board of the University of Guelph who made this research possible through generous financial support.

Special gratitude is extended to C. P. Alderfer, Yale University, for most valuable comments received on earlier theoretical aspects of the application of the E.R.G. theory to the consumer role, and for reviewing the manuscript of the present work.

Requests for reprints should be sent to Gerhard W. Scherf, Department of Consumer Studies, University of Guelph, Guelph, Ontario, Canada.

The general possibility of using Proposition 2 of the E.R.G. theory to explain some of the problems of consumer dissatisfaction has already been advanced elsewhere (Scherf, 1974). The purpose of the present study is to test empirically several hypotheses which can be deducted from such a frame of reference. The following hypothesis is derived from the E.R.G. theory's Proposition 2:

Hypothesis 1. The less consumers' relatedness needs are satisfied, the more existence needs in the form of income and consumer goods they desire.

Low levels of relatedness satisfaction would eventually be expected to cause low levels of existence satisfaction, due to the consumer's inability to satisfy the new existence needs. Such a proposition is not readily assessed in an unbiased way by means of the present study. This project could, however, be extended to explore possible causal relationships.

On the other hand, it is very likely that an income differentiation will exist with regard to the extent to which consumers *can satisfy* additionally created existence needs: It should be easier to meet them with higher incomes than with lower incomes (although the actual level of income does not necessarily affect consumers' desire for more existence needs). The following hypothesis is therefore advanced:

Hypothesis 2. As the consumers' capacity to meet additional existence needs increases due to higher actual incomes, the rating of satisfaction with existence needs increases accordingly.

Finally, it is proposed to verify the applicability of Proposition 1 of the E.R.G. theory to man in the consumer role. Proposition 1 of the E.R.G. theory states that the less existence needs are satisfied, the more they will be desired. The hypothesis for the present context is formulated as follows:

Hypothesis 3. The less consumers' existence needs in the form of income and consumer goods are satisfied, the more they are desired.

This application of Proposition 1 is considered especially important in view of the fact that the approach used in the present study for the assessment of existence needs for man as a consumer differs significantly from that used previously for studies of man in an organizational setting.

In view of the hierarchically different positions occupied by existence and relatedness needs, it is necessary, furthermore, to consider the possibility that a person's actual income not only governs his satisfaction with existence needs as postulated in Hypothesis 2 but also affects the fundamental relationships between existence and relatedness needs. Especially for the lowest income group it is possible that the deprivation of essential existence needs is so intense that the relatedness needs are hardly yet considered. This problem is examined by selectively excluding the poorest segment of the sample in a separate analysis. Similarly, it is considered necessary to examine whether age factors might also place limits on the applicability of the theory.

METHOD

Subjects and Procedure

The subjects, approximately 306 consumers, were drawn from a small city and one of its neighboring towns in the province of Ontario. Respondents for the town population were selected by means of a random process using the name lists of the Regional Assessment Office. They were contacted personally by the research assistant, who provided them with the questionnaire, and asked that it be returned by

TABLE 1

MEAN RATINGS OF CONSUMERS' DESIRE FOR MORE EXISTENCE NEEDS (D–E) AT VARYING LEVELS OF SATISFACTION WITH RELATEDNESS NEEDS (S–R)

Study	S–R Rating						Analysis	
	1	2	3	4	5	6	*F* test	η^2
1. Total sample								
n	89	154	45	7	2	2		
M	5.64	5.97	5.58	6.57	8.00	7.50	ns	.03
SD	1.70	1.70	1.61	1.59	0	1.50		
2. Age groups [illegible]5 only								
n	78	140	40	5	2	1		
M	5.64	6.04	5.55	6.50	8.00	9.00	$p < .05$	.05
SD	1.75	1.65	1.59	1.36	0	0		
3. Age groups [illegible]0 only								
n	62	120	31	4	2	1		
M	5.55	6.21	5.61	6.25	8.00	9.00	$p < .01$	.11
SD	1.51	1.60	1.65	1.09	0	0		
4. Income groups above $5,000 only								
n	59	100	30	5	2	2		
M	5.58	5.99	5.60	6.20	8.00	7.50	$p < .01$	.08
SD	1.70	1.74	1.56	1.42	0	1.50		
5. Total sample								
n	27	116	64	37	11	3		
M	5.22	5.69	5.84	6.05	7.00	7.33	$p < .05$	.05
SD	1.62	1.69	1.38	1.86	1.81	1.25		

Note. Studies 1–4 relate satisfaction with relatedness needs with significant others; Study 5 relates satisfaction with relatedness needs with people at the place of employment.

mail. A total of 59 consumers completed the questionnaire satisfactorily (61% of the questionnaires distributed). Respondents for the city population were selected by randomly identifying small districts within which every family was then contacted by the research assistant. The assistant handed the questionnaire to the respondent and asked that it be ready for personal collection by a specific time the following day. A total of 247 consumers in this group answered the questionnaire satisfactorily (73% of the questionnaires made available). Homogeneity tests (*F* tests) were carried out with respect to the response patterns of both the existence needs ratings and the relatedness needs ratings of the two groups before they were combined into a single sample.

Every consumer was given assurance of complete anonymity both verbally and in writing on a cover letter which accompanied the questionnaire. The only information regarding the project given to the subjects was that its purpose was to assist the Department of Consumer Studies "in finding possible solutions to many of the problems which people face in today's markets and in our society in general."

Measures

Alderfer (1969) assessed man's existence needs in the role of producer (employee) primarily through questions which related to the strength of desire for, and satisfaction with, income and fringe benefits. Man in his role of consumer, however, has principally two alternatives for expressing his existence needs: He may perceive his level of consumption to depend on his income and/or on his fate in the marketplace. The desire for existence (D–E) need was therefore measured by giving equal weight to the rating of the respondent's desire for more income from employment sources and to the rating of the desire for "more" in terms of wants related to the marketplace: $\text{D–E} = \text{D}_{\text{Income}} + \text{D}_{\text{Market}}$. While the desire for more income could be assessed in one direct question, the desire for more consumption opportunities from the marketplace was measured as the average response to five questions: The desire for (*a*) better prices, (*b*) better quality, (*c*) better information, (*d*) the ability to replace outstyled goods more frequently, and (*e*) the ability to replace worn-out goods more frequently. The data concerning the internal consistency of the components of the multiple scale are explored in detail in a separate article (Scherf, 1973).

As in Alderfer's work, all satisfaction measures were based on 6-point Likert-scale items, while the need intensity measures were ratings of desire for (*a*) no more, (*b*) slightly more, (*c*) somewhat more, (*d*) much more, and (*e*) very much more, of the selected need category. The numerical values, 1–6, assigned to the satisfaction ratings, likewise express greater dissatisfaction with increasing numbers.

Only two categories of relatedness needs were explored: (*a*) with "significant others," which included close friends, family members, and spouse, and (*b*) with people at the place of employment, which included co-workers and supervisors.

Although it might appear that persons such as sales representatives or other store employees who serve in direct contact with the consumer could be regarded as falling into the category of significant others, they have not been included in the present survey. Their exclusion was justified on the basis that those included in significant others are a part of a relatively durable system of interaction for a person in his consumer role, while the contact with sales personnel, in contrast, is rather transitory by nature and relatively brief in duration, and hence is unlikely to have a pronounced influence on a consumer's desire for "more" or his satisfaction with existence needs. It was felt that the salesman is more likely to influence a consumer's satisfaction with a very specific piece of merchandise rather than his perception of consumption satisfaction in general.

RESULTS

The data used for testing Hypothesis 1 are summarized in Table 1. The respondents were stratified according to their rating of satisfaction with relatedness needs in two different contexts, satisfaction with relatedness needs with significant others (Studies 1–4) and satisfaction with relatedness needs with people at the place of employment (Study 5), and the mean ratings of each group's desire for more existence (D–E) needs were calculated. In every study the D–E means increased quite systematically as the satisfaction with relatedness needs decreased (increasing satisfaction with relatedness ratings), in apparent agreement with the hypothesis. Particularly noteworthy is the fact that the last three groups, Studies 2, 3, and 4, that is, those groups reflecting "disagreement" responses on the Likert scale, contained very few respondents, but invariably showed higher means for desire for more existence needs than did those of the first three groups, that is, those reflecting an "agreement" response on the Likert scale. However, the *F* test performed for each series of means showed that for the total sample statistical significance at the .05 level was obtained only in Study 5 for satisfaction with relatedness needs with people at the place of employment.

Studies 2–4 were carried out in order to establish whether qualifications regarding age and income would affect the data sought for the support of Hypothesis 1. Accordingly, the

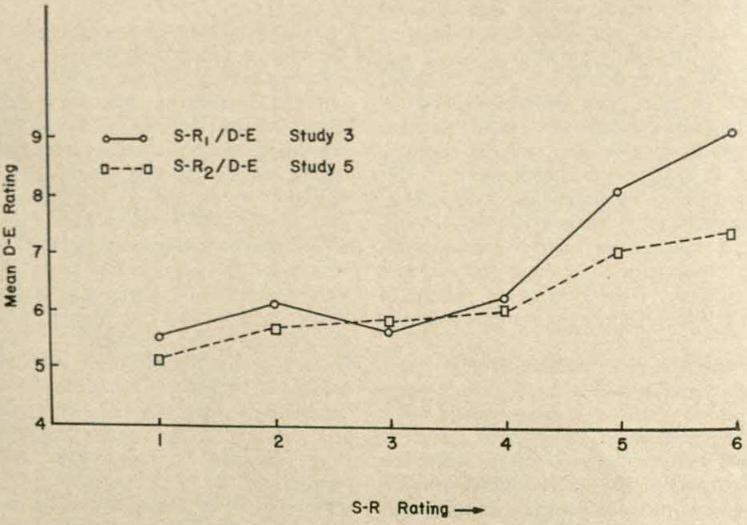


FIGURE 1. Graphical representation of the data of Studies 3 and 5.

TABLE 2

MEAN RATINGS OF CONSUMERS' SATISFACTION WITH EXISTENCE NEEDS (S-E) IN VARYING INCOME CLASSES IN STUDY 6

Measure	Income group				Analysis	
	Below $5,000	$5,000–10,000	$10,000–15,000	Above $15,000	F test	η^2
n	90	109	43	42		
M	7.70	7.71	6.19	5.38	$p < .01$	.16
SD	2.29	2.08	2.16	1.09		

means analysis of the satisfaction with relatedness needs with significant others stratification used in Study 1 was repeated after selectively eliminating those groups of respondents who theoretically may be considered as exceptions to the assumed need hierarchy. The data of Studies 2 and 3 indicated that the statistical significance expressed by the F test reached the .05 level when respondents between the ages of 20 and 65 only were considered, and it improved to the .01 level when the age groups included only those between 20 and 50. A similar improvement was obtained in Study 4 when the respondents of the lowest income group were eliminated from the sample.

A closer inspection of the data suggested that the association between the two variables, satisfaction with relatedness needs and desire for existence needs, may be of a curvilinear rather than linear nature. A graphical illustration of this trend is shown in Figure 1 for two selected studies. In order to obtain an adequate measure of the strength of the total association between both variables, the corresponding correlation ratios, η^2 (Hays, 1963) were calculated. The η^2 values confirmed the conclusions drawn from the means/F tests regarding the effect of age and income on the satisfaction with relatedness needs/desire for existence needs relationship.

For those studies where the F test indicated that the differences among the means were statistically significant, yet the total association revealed by the η^2 reached only relatively small values, it was of further interest to determine *where* the main differences would occur. Multiple t comparisons were therefore carried out for Studies 2–5 of Table 1.[2] It appeared that the significant differences ($p < .05$) occurred almost exclusively when the comparison involved one group whose satisfaction with relatedness needs rating was derived from an "agree" response of the Likert scale, and one whose satisfaction with relatedness needs rating was derived from a "disagree" response of the Likert scale.

Study 6 was carried out for the examination of Hypothesis 2. Respondents were stratified according to their reported income group, and the mean response to their satisfaction with existence needs was calculated. The data, recorded in Table 2, agreed with the hypothesis: The satisfaction with existence needs increased systematically with higher income levels. The corresponding graphical presentation of the data, shown in Figure 2, suggested the existence of a curvilinear relationship between these variables.

Hypothesis 3 was examined in Study 7. Respondents were stratified according to their desire for existence needs rating, and the mean responses for satisfaction with existence needs were calculated. The data, recorded in Table 3, supported the hypothesis: The desire for existence needs increased as the satisfaction with existence needs decreased. A closer inspection of these data

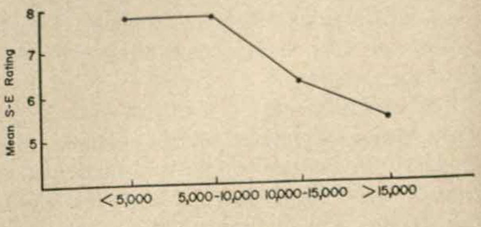


FIGURE 2. Graphical representation of the data of Study 6.

---

[2] Detailed results are available from the author upon request.

TABLE 3

RESPONDENTS' MEAN RATINGS OF SATISFACTION WITH EXISTENCE NEEDS (S–E) AT VARYING LEVELS OF DESIRE FOR MORE EXISTENCE NEEDS (D–E) IN STUDY 7

Measure	D–E Rating									Analysis	
	2	3	4	5	6	7	8	9	10	F test	η^2
n	4	16	47	61	61	57	22	21	4		
M	4.00	4.94	5.66	7.26	6.98	8.07	8.95	9.14	9.50	.01	.36
SD	0	1.03	1.77	1.91	1.32	1.66	1.65	2.09	1.06		

suggested that the trend towards curvilinearity may be less pronounced for the desire for existence needs/satisfaction with existence needs association than for the satisfaction with relatedness needs/desire for existence needs associations discussed previously.

DISCUSSION

The finding that Proposition 2 of the E.R.G. theory is applicable to man in the consumer role, although with some limitations as to income and age, is perhaps the most important result of the study. The fact that the more relatedness needs are frustrated, the more existence needs will be desired may add an entirely new dimension of strategy to those concerned with making consumers more satisfied in today's environment. It would appear that real improvement of satisfaction could occur only if man's meeting of his relatedness needs is facilitated along with the improvement of his economic welfare.

In accordance with most need hierarchies, the most elementary and vital existence needs have to be reasonably satisfied before higher needs such as relatedness needs, take on major importance; hence it is not too surprising to find that the interdependency of desires for more existence needs and interpersonal satisfactions was more pronounced when persons of very low incomes were excluded from the sample.

The exclusion of the youngest and of the most senior segments of the sample also produced an increase of the satisfaction with relatedness needs/desire for existence needs relationship. This finding might be tentatively explained by assuming that to persons of a very young age, the opportunity to translate frustrated relatedness needs into additional existence needs is not yet available because of their still limited earning power. On the other end of the scale, persons of the most senior segment of the sample may once again find themselves in the same restricted financial position; alternatively, of course, it is possible that senior persons, because of their life experience, no longer consider the regressive transfer of frustrated relatedness needs into additional existence needs. (Concerning this group, future studies which are designed to assess the effect of frustrated relatedness needs on growth needs will be of great interest).

It is acknowledged that the index used for measuring existence needs in the consumer role may still be incomplete. Only five different items, for example, were chosen to express the strength of desire for "better deals" in the marketplace, while a much larger number would be needed if one wished to exhaust all possible alternatives.

The combination of measuring satisfaction items on a 6-point Likert scale and desire items on a 5-point scale has been adopted without renewed qualification from the original work on the E.R.G. theory (Alderfer, 1969). As all the items were phrased so that satisfaction was at the 1 end and the desire scales had high ratings on the 5 end, the possibility exists that some responses were biased by an individual's mere preference for high numbers.

Furthermore, the present work shares with Alderfer's work the limitation that both are correlational studies and hence do not permit causal inference.

[illegible]EFERENCES

Alderfer, C. P. [illegible]mpirical test of a new theory of human needs. [illegible]ganizational Behavior and Human Perform[illegible] 1969, 4(2), 142–174.

Hays, W. L. *St[illegible]s for psychologists.* New York: Holt, Rineha[illegible] Winston, 1963.

Hulin, C. H. [illegible]ources of variation in job and life satisfacti[illegible]he role of community and job-related varia[illegible] *Journal of Applied Psychology,* 1969, **53**, 27[illegible].

Scherf, G. W. H. Towards an appreciation of the sources of consumers' dissatisfaction: Correlations between alternative expressions of the desire for more consumption opportunities. Unpublished manuscript, University of Guelph, 1973.

Scherf, G. W. H. Consumer education as a means of alleviating dissatisfaction. *Journal of Consumer Affairs,* 1974, **8**, 61–75.

(Received June 11, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 4, 472–482

SEAT BELTS:
RELATIONS BETWEEN BELIEFS, ATTITUDE, AND USE[1]

GUNILLA FHANÉR[2] AND MONICA HANE
University of Uppsala, Sweden

A sample of 368 car owners answered a mailed questionnaire constructed on the basis of a model of attitudes toward seat belt use. A factor analysis yielded a belief pattern interpreted in terms of five factors, labeled "discomfort," "worry," "risk," "effect," and "inconvenience." The model appeared useful since an independent measure of attitude to seat belt use (A_o) could be predicted ($r = .84$) from a linear combination of these factor scores. The correlation between the attitude measure and reported use of seat belts was .56, or about the same as that between the belief factor combination and reported use. The combination of the "discomfort" and "effect" factors gave near optimal predictions ($r = .52$). The results were replicated on a new sample. A model of seat belt use was suggested in which conceptions about discomfort of belt usage and effects of belts in an accident were regarded as determinants of usage.

The present study is part of a project aiming at investigating variables relevant for the use of seat belts in private automobiles. Knowledge of such variables should be of value in designing campaigns with the purpose of raising the level of seat belt usage by motorists.

In a previous study it was shown that individual variation was responsible for a larger proportion of belt use behavior variance than was situational variation (Fhanér & Hane, 1973b). This finding led to a search for individual variables related to belt usage in the hope of finding variables which could be influenced by campaign measures.

As was pointed out in a literature survey (Fhanér & Hane, 1973a) the striking fact is that belt users and nonusers alike express a very favorable attitude to seat belts, the belt users being only somewhat more favorable. However, a serious limitation of practically all studies so far is that only one item at a time has been used as an indication of attitude, probably resulting in low reliabilities and, hence, low correlations with usage.

The popularity and extensive use of the attitude concept has led to low conceptual precision. As early as 1935 Gordon Allport (1967) considered more than 100 different definitions in his review of the general area of attitude theory and research (Allport, 1967; see also McGuire, 1969). In the present study the concept of attitude is regarded as a latent variable referring to "the amount of affect for or against a psychological object [Thurstone, 1931]." Cognitive and action variables are seen as determinants or consequents of attitude.

The attitude model adopted in this study is consistent with that proposed by Rosenberg (1956, 1960) within a cognitive consistency theoretical framework, and arrived at by Fishbein from a behavior theory point of view (Fishbein, 1967a). Rosenberg predicted that the affect attached to an attitude object will be highly related to (*a*) "the perceived instrumentality of the attitude object," that is, the judged belief or probability that the attitude object would lead to, or block, the attainment of "valued states," and (*b*) the "value importance," that is, the amount of affect expected from these "valued states." Similarly, Fishbein stated that an individual's attitude toward any object is a function of (*a*) the strength of his beliefs about the object and (*b*) the evaluative aspects of those beliefs (Fishbein, 1967a). This was expressed as follows:

---

[1] This study is part of a research project concerning factors influencing the use of seat belts in private automobiles. The project was carried out by the authors under the Swedish Road Safety Office Contract Inf. 32, and it was supported by a grant from the Swedish Renault, Inc. (AB Svenska Renault). The complete report, No. 25, Dnr 93-732109 u, is available at the Swedish Road Safety Office, Driver's License and Research Departments, Fack, S-171 20 Solna, Sweden.

[2] The authors are equally responsible for the investigation and its presentation.

Requests for reprints should be sent to Gunilla Fhanér, Kristina Hards vag 9, S-443 00, Lerum, Sweden.

$$A_o = \sum_{i=1}^{N} B_i a_i$$

where

A_o = the attitude toward object o,
B_i = the strength of belief i about o, that is, the "probability" or "improbability" that o is associated with some other concept x_i;
a_i = the evaluative aspect of B_i, that is, the evaluation of x_i;
N = the number of beliefs about o.

An individual's attitude regarding seat belt use can thus be seen as a function of the algebraic sum of the products obtained by multiplying the strength of his beliefs associated with belt usage by the evaluative aspects of those beliefs.

An extensive literature search, formal and informal discussions with motorists of many categories, experts and nonexperts, as well as some preliminary studies based on questionnaires and interviews at a gas station yielded a variety of beliefs related to seat belt usage.

The injury preventive effect of seat belts seemed to be a salient belief of most people. However, concern about seat belts aggravating injuries was also frequently expressed. For instance, some people expressed the belief that a seat belt might be difficult to unlock in the event that the car caught fire or landed in the water.

A feeling of constraint and discomfort seemed to be rather widespread, and was a common excuse for nonuse. For some people this feeling seemed to be a general and deep-rooted aversion against being tied up, but for others it appeared to concern the restricted area of motion, the difficulty to reach, etc. On the other hand, some people felt increased comfort from wearing a seat belt, either by the physical support offered by a belt, especially on curbed roads, or by the generation of a feeling of security.

It was also held that the use of seat belts could influence the risk of accidents, either by making people feel too secure, leading to careless driving, or by making drivers less anxious, leading to safer driving.

The inconvenience aspect of belt usage was frequently mentioned. Thus, many remarks were made concerning difficulty to lock and unlock and to adjust to fit.

The image of a driver also seemed to be affected by belt use. To some people seat belt usage was a sign of being a good and careful driver and of being responsible and having good judgement, while to some others a seat belt user was an anxious and insecure driver.

Using Fishbein's (1967a) model, the attitude to seat belt use (A_o) for an individual can be estimated from a simple regression equation involving these beliefs related to seat belt usage.

$$A_o = B_1 a_1 + \ldots + B_5 a_5$$

We shall make the simplifying assumption that each evaluative aspect, a_i, has the same value for all individuals. The variation in A_o can thus be predicted from variation in each B_i or belief strength.

According to Fishbein (1967b) and Fishbein and Raven (1967) the strength of a belief can be measured in probability terms. The strength of the belief concerning the relation between seat belt use and injuries can be regarded as a conditional probability. First, there is the probability of an accident, and second, the probability of a seat belt reducing or aggravating injuries, given an accident. The strength of this belief could thus be measured by a two step process, whereby the perceived accident risk is first obtained and then combined with the belief about seat belt effects, given the accident.

As a strategy for finding variables causally related to belt usage, it would be wise to start by investigating existing correlations between belt usage, attitude, and the beliefs discussed previously.

STUDY 1

The primary aim of the present investigation was the identification of attitudinal and cognitive variables related to the use of seat belts. A subordinate goal was the construction of an instrument, tapping important beliefs, which could be used in further studies.

Method

Measures

Attitude to belt use (A_o). To get a "direct" measure of attitude to belt usage (A_o) seven statements intended to tap the evaluative aspect of belt use were constructed, such as "Everyone ought to use a seat belt when driving a car." The respondents indicated on a 5-step verbal scale their degree of agreement with each statement: completely agree (1), mainly agree (2), partly agree (3), mainly disagree (4), and completely disagree (5). "No opinion" could also be checked.

Beliefs. On the basis of the beliefs about seat belts and belt usage that seemed most salient, according to the previous discussion, 52 Likert-type items were constructed. The respondents indicated their degree of agreement by using the aforementioned 5-step verbal scale. The content of the items is given in Table 1.

Reported seat belt use. To minimize possible influence of behavior reports on the answers to the belief and attitude items information on the respondent's actual seat belt usage was obtained two weeks after the questionnaire had been returned. The degree of influence of the questionnaire responses on the behavior reports was assumed to be less than in the reversed direction (see Wicker, 1969). The respondents were contacted by telephone and were asked to indicate their habitual seat belt usage in city traffic and on the highway. The responses were coded in four categories: never or almost never (1), sometimes or seldom (2), often or generally (3), and always or almost always (4). According to the one dimensional model presented in a previous study (Fhanér & Hane, 1973b), measures of seat belt usage in city and highway driving can be summated to an index of overall belt usage. Level of belt usage is thus given in terms of a 7-step scale, a value of 2 indicating nonuse and of 8 indicating consistent use in both highway and city driving. Those lacking seat belts in their cars (10 respondents) were coded as nonusers. The reliability of self-reports of belt usage in terms of homogeneity has been shown to approach unity (Fhanér & Hane, 1973b), and the validity of such reports has been demonstrated to be relatively good (Fhanér & Hane, 1972, 1973b). In the following, "seat belt use" or "belt usage" refers to the summated, 7-step scale.

Background variables. Beside the belief and attitude items the questionnaire also contained questions concerning type and age of car, type of belt, age of driver's license, mileage driven each year, sex, level of formal education, and size of home community.

Subjects

From the automobile registration files in Örebro county (län) covering 80,000 cars, a sample of 600 car owners was randomly drawn. After the exclusion of business firms, official authorities, etc., the sample consisted of 545 private passenger car owners. Örebro county is located in the south central parts of Sweden and is often referred to as a "Miniature Sweden," since its demographical composition quite closely resembles that of Sweden.

Procedure

The questionnaire was mailed to the respondents on December 8, 1970. Enclosed was an introductory letter pointing out the scientific purpose of the study and stressing the inaccessibility of an individual's answers to people outside the project. To increase the motivation to respond a lottery was arranged among those who returned their complete questionnaires. Complete information was obtained for 368 subjects (68%).

One third of the nonrespondents were contacted by telephone. Thirty-two were reached, but eight refused to answer any questions. Those who responded showed roughly the same distribution of reported belt usage as did the respondents. Population statistics for car owners in Örebro county were available for the age and sex variables (Statistiska Centralbryån, 1970). In these two variables the distribution of the present sample was almost identical to that of the population. Thus, even though subject loss probably is never wholly random, no reason was found for a suspicion of biased results due to subject loss.

Results

The index of seat belt use and the background variables were cross-tabulated. Sex was the only variable that was significantly related to belt usage, women reporting a higher level of usage than men ($\chi^2 = 18.40$, $df = 2$, $p < .001$).

Attitude to Seat Belt Use (A_o)

The intercorrelations between the seven items ranged from .25 to .62. The scores on these items were added for each individual to get a measure of his attitude (A_o). The small number of respondents (mean 4.1%) who had checked the "no opinion" alternative was included in the middle response category 3, "partly agree." The reliability of the attitude measure in terms of homogeneity was estimated with the Kuder-Richardson (K-R) 20 formula and was found to be .84.

Beliefs about Seat Belt Usage

Factor analysis. To investigate whether the hypothesized belief structure with respect to seat belts and the use of them was an

TABLE 1

ROTATED FACTOR MATRIX

Item and content	Factor					Communality
	I	II	III	IV	V	
242 . . . feel confined	.80	—	—	—	—	.73
174 . . . feel constraining	.67	—	—	—	−.32	.72
159 . . . makes one irritated . . . worse driver	.64	—	—	—	−.20	.65
155 . . . feel suffocated	.56	—	—	—	—	.37
141 . . . sits more comfortably	−.48	—	.30	—	—	.37
133 . . . causes accidents	.48	—	—	−.23	—	.44
131 . . . a sign of overcautiousness	.44	—	.24	−.29	—	.48
228 . . . always uses . . . more nervous	.42	—	—	—	−.22	.44
247 . . . view of the road . . . difficult	.39	—	—	—	.32	.45
140 . . . difficult to reach	.38	—	—	—	−.37	.37
210 . . . false sense of security	.35	—	—	−.24	—	.37
164 . . . insecure drivers use	.32	—	.20	−.29	—	.44
125 . . . often worry . . . accident	—	.81	—	—	—	.66
270 Before driving . . . often worry	—	.64	—	—	—	.42
151 . . . can not help worrying	—	.59	—	—	—	.37
238 . . . always worry . . . family is driving	—	.58	—	—	−.22	.39
268 . . . seated . . . never think of risk	—	−.49	.27	—	—	.33
128 During winter . . . awful to drive	—	.47	—	—	—	.24
156 . . . never think of being injured	—	−.41	.27	—	—	.24
130 . . . anything can happen . . . ?	—	−.40	—	—	−.23	.22
173 . . . never think of accidents	—	−.49	.50	—	—	.46
230 . . . easy to resist . . . by the arms	—	—	.47	—	—	.39
171 . . . possibilities to resist . . . by his arms	—	—	.45	—	—	.28
211 . . . supports the body	−.40	—	.44	—	—	.34
169 . . . feelings of security . . . drives safely	−.43	—	.43	—	—	.47
260 People exaggerate the risks	—	—	.41	—	—	.26
165 . . . accidents avoided . . . by good judgment	—	—	.40	—	—	.16
142 . . . car is just as safe as	—	−.27	.34	—	—	.20
143 . . . certain group in accidents	—	—	.34	.21	—	.16
135 . . . driving carefully . . . avoids accidents	—	—	.33	—	—	.12
267 . . . a group of accident prone drivers	—	—	.33	—	—	.18
258 . . . reminds of risk . . . drive carefully	—	.29	.31	—	—	.22
246 . . . escape injury by getting down	.20	—	.31	—	—	.23
261 It has been proven . . . escape injuries	—	—	—	.66	—	.55
265 . . . by preventing being thrust out	—	—	—	.60	—	.38
231 . . . prevent casualties	—	—	—	.54	—	.26
245 . . . effective . . . against brain damage	—	—	—	.51	—	.31
249 . . . highway . . . no effect	—	—	—	−.43	−.21	.44
136 It has not been proven . . . prevent injuries	—	—	.28	−.41	—	.27
226 Only lighter injuries prevented	—	—	.30	−.39	—	.43
162 . . . better be thrown clear out	.21	—	.21	−.34	—	.30
147 . . . satisfactory protection	—	—	.27	.32	—	.25
227 . . . trouble putting on	—	—	—	—	−.83	.71
163 . . . rather simple to put on	—	—	—	—	.81	.63
257 . . . trouble to adjust	—	—	—	—	−.72	.48
158 . . . gets caught in the door	.26	—	—	—	−.52	.43
150 . . . wrinkle the clothes	.27	—	—	—	−.38	.30
229 . . . difficult to undone . . . if one fire	.29	—	—	—	−.36	.43
145 One out of three . . . injured	—	.22	—	—	—	.09
148 Even . . . experienced drivers . . . involved	(.16)	—	—	—	—	.05
160 . . . won't drive an inch . . . cautious	.29	—	—	—	—	.18
209 . . . side ways . . . disadvantage to wear	—	—	—	−.26	—	.23
241 . . . into the water . . . not get unlocked	—	—	.26	−.22	−.28	.26
256 . . . side collisions . . . no help	—	—	—	−.28	−.25	.27

adequate description of that of our respondents, the product-moment correlations among the items attempting to tap such beliefs were factor analyzed. A principal-factors solution with squared multiple correlations as communality estimates was used. Oblique rotation was performed with the rotation criterion for simple loadings suggested by Jennrich and Sampson (1966). Before the rotation "Kaiser normalization" of the communalities was performed. Five factors emerged that had eigenvalues greater than 1.0. These five factors accounted for about 36% of the total variance, and the extraction of additional factors contributed very little to that proportion. Of the 54 variables 41 had only one loading above .30, and no variable had more than two such loadings. The rotated matrix and communalities are presented in Table 1. The items are arranged in accordance with the factor pattern to facilitate interpretation of the factors. Loadings below an arbitrary limit of .20 have been omitted to increase the readability of the matrix.

The intercorrelation between Factors I and V was −.42, between I and IV −.36, between I and III, as well as IV and V, just above .20. The remaining intercorrelations were close to zero.

Interpretation of the factors. The items "I feel uncomfortly confined wearing a seat belt" and "The seat belt feels constraining, I think" had the highest loadings in Factor I and low loadings in the other factors. The largest negative loading in the first factor was received by the item "One sits more comfortable wearing a seat belt." Thus, Factor I seems to cover feelings of discomfort.

The second factor had its highest loadings in items expressing worry, for example, "I often worry about getting involved in a car accident." Large negative loadings in this factor were typical of items denying thoughts about accidents and injuries, exemplified by "I never think of the possibility of being seriously injured in a car accident." An appropriate label for Factor II appears to be "worry."

The items with rather pure loadings in Factor III concerned the low risk of getting into accidents and of being injured without wearing a seat belt. Hence, Factor III seems to represent feelings of risk.

The fourth factor can easily be interpreted in terms of seat belt effectiveness. "It has been proven that the use of a seat belt is one of the best ways to avoid injuries in a collision" had the highest loading, and the negation of that statement had a large negative loading. Factor IV will accordingly be labeled an "effect" factor.

The interpretation of Factor V also seems rather straightforward. "It is a lot of trouble putting on and taking off a seat belt," and the obverse "Seat belts are rather simple to put on and take off" had the largest absolute values. Thus, inconvenience in handling seems to be the common denominator of Factor V.

Three of the hypothesized factors can easily be identified, namely the Effect factor (IV), the Discomfort factor (I), and the Inconvenience factor (V). The items constructed to cover the experience of risk were evidently not homogeneous but seemed to cover two dimensions not closely related. Possibly Factor II, Worry, rather represents a personality trait such as anxiety, while Factor III, Risk, appears to have a more cognitive character. The items concerning the image of a belt user and the items on seat belts increasing the accident probability did not form separate factors, but had moderate loadings in several factors.

Individual factor scores. The most simple way of estimating a factor score for an individual is to select the variables loaded only in that factor, and to sum the scores of those variables. The variables are given equal weights thus enabling the estimation of a factor score in new studies with different individuals without the computation of a new factor analysis (Sjöberg, 1972). Four marker variables were chosen for each factor. The agreement between the matrices of intercorrelations between factors and factor scores, respectively, was quite close. The reliabilities of the factor scores in terms of homogeneity were computed using the K-R20 formula and ranged from .86 for Factor I, Discomfort, to .55 for Factor III, Risk.

Since it could be suspected that not only the reliabilities of the factors but also the

TABLE 2

Beliefs Correlated with A_o and with Reported Use

Belief factors	A_o		Reported use	
	Sample 1 ($n = 184$)	Cross-validation, Sample 2 ($n = 184$)	Sample 1 ($n = 184$)	Cross-validation, Sample 2 ($n = 184$)
I. Discomfort (D)	−.80	−.72	−.53	−.49
II. Worry (W)	.14	.01	.16	.12
III. Risk (R)	.30	.31	−.02	.13
IV. Effect (E)	.44	.53	.24	.28
V. Inconvenience (I)	−.47	−.41	−.44	−.33
D + E	.82	.77	.53	.51
D + I	.81	.72	.57	.49
D + E + I	.83	.77	.57	.50
D + R + E + I	.83	.78	.60	.48
D + W + E + I	.83	.77	.59	.52
D + W + R + E + I	.84	.78	.61	.50

Note. A_o = attitude to seat belt use.

belief pattern varied among individuals with different levels of belt usage, separate factor analyses of the 20 factor score variables were performed for the constant users ($n = 62$) and nonusers ($n = 96$), respectively. Both analyses yielded a pattern that was similar to the original pattern. Inspection of the communalities revealed no systematic differences between the two extreme groups. It was, thus, concluded that neither the pattern of beliefs, nor the reliabilities of the factor scores were dependent upon level of belt usage.

Correlation between Attitude to Seat Belt Use, A_o, *and Reported Use*

A cross-tabulation between the attitude measure, A_o, and the index of belt usage indicated that the regression was approximately linear. Thus, the product-moment correlation coefficient of .56 would be a fair estimate of the relation. Due to imperfect reliabilities of the variables it is of course an underestimation of the "true relationship." The results are interesting in light of the many studies showing a lack of relationship between seat belt use and attitudes to belts and belt usage (Fhanér & Hane, 1973a).

Prediction of A_o *from Beliefs about Seat Belt Usage*

To investigate how well the attitude to belt usage could be predicted from beliefs in terms of factor scores, multiple correlations were calculated for one half of the sample ($n_1 = 184$). The regression equations thus obtained were then used for the other half ($n_2 = 184$), and the product-moment correlations between the predicted and obtained values were computed. The predictions were made on the basis of single factors as well as on those factor combinations that seemed most promising. The results can be seen in Table 2.

The most striking result presented in Table 2 was the high correlation between Factor I, Discomfort, and attitude to seat belt use, A_o. Combining Factor I with one or more factors barely improved on the predictions for Sample 1, although it tended to reduce the shrinkage in Sample 2 somewhat. Maximal values of the multiple correlation coefficient were obtained with a linear combination of all five factors, but summing the scores of Factor I, Discomfort, and Factor IV, Effect, gave a very good approximation. The coefficients for the total sample were .82 and .80, respectively.

Prediction of Reported Belt Use from Beliefs about Usage

As in the prediction of seat belt attitude, A_o, correlations between beliefs and the belt use index were calculated on Sample 1 and compared to those achieved for Sample 2, using the regression equations obtained for Sample 1. The results are presented in Table 2.

Factor I, Discomfort, was the best single predictor of belt usage. Factor V, Inconvenience, rated second, but the addition of Factor V to Factor I did not improve on the prediction due to a high intercorrelation between those two factors. Factor II, Worry, and Factor III, Risk, had both very weak relations to usage, while Factor IV, Effect, and usage had a somewhat higher correlation.

The regression equation that yielded the best prediction after shrinkage included Factors I, II, IV, and V. To obtain optimal estimates of the regression coefficients they were recalculated for the total sample ($n = 368$) yielding a multiple correlation of .56, when the four factors were combined, and .52 for the Discomfort–Effect combination. On the whole, reported belt usage could be predicted equally well from beliefs in terms of factor scores as from the attitude index, A_o. This should be expected on the basis of the high relation between beliefs and A_o.

As was stated in the introduction the belief, B_1, that is, the instrumentality of seat belts with respect to injuries, was conceptualized in terms of a conditional probability. The probability of a car accident should, hence, be combined with the probability of a seat belt reducing or aggravating injuries, given an accident. This line of thinking implied that a conjunctive model for weighting the accident probability and the Effect would give better predictions than a linear model. Predictions based on a conjunctive model were thus computed and were compared to those based on a linear model. Since it was not obvious whether the Worry or Risk factor best reflected the accident probability, Effect was combined with both factors. However, using the conjunctive model did not improve on the predictions.

STUDY 2

The aims of the present study were twofold: first, to determine the generality of the seat belt belief pattern obtained in Study 1 across individuals; second, to check the measurement characteristics of the factor instrument, particularly its predictive validity in terms of self-reports of behavior as well as behavior observations.

Method

Measures

Beliefs. A questionnaire was constructed on the basis of the results in Study 1. It consisted of the four marker variables on each of the Factors I, Discomfort, II, Worry, IV, Effect, and V, Inconvenience that had yielded the best prediction of reported belt usage.[3] The responses were coded from 1 to 5 as in Study 1.

Reported seat belt use. The first part of the questionnaire contained 17 questions on belt usage in various situations (see Fhanér & Hane, 1973b). Included, also, were the items on frequency of usage in city traffic and in highway traffic, which were used to form the seat belt use index in Study 1. As in Study 1 a 7-step belt use index was constructed by summation of the scores on the items concerning frequency of usage in city and in highway driving.

Observed seat belt use. The observations were carried out in the rush hours on [illegible] workday mornings chosen at random from November 17, 1970 to February 3, 1971. The observer was hidden in a worker's hut at a road intersection where around 300 cars passed every morning. License plate numbers and seat belt usage of the drivers were read into a tape recorder. Thus, for each car the frequency of passage and of driver's seat belt usage could be obtained. A reliability check of the observations showed that two independent observers reported the same number of belt users but that only 73% coincided. The reliability of the observations was thus not perfect. Especially single registrations of seat belt use in a series of nonuse for an individual should be regarded with caution (see Fhanér & Hane, 1973b).

Subjects

The sample consisted of 105 drivers. Of these, 25 were defined as seat belt users as they had been observed wearing a seat belt at least once out of a minimum of five observations. The consistent 80 nonusers were randomly drawn among those who had been observed at least eight mornings. The mean number of available observations exceeded 8 for the "users" and 10 for those defined as "nonusers." The average proportion of seat belt use instances for the users was .61. Thus, there was no basis for assuming that the present sample was representative of drivers in general. However, it turned out to be highly similar to the sample in Study 1 with respect to background variables. The subjects were identified through the registration files. They

---

[3] As an exception Item 141, with the fifth highest loading in Factor I, was included instead of Item 159 ranking third. The choice was based on a preliminary analysis which had resulted in a higher loading for Item 141 than for Item 159.

TABLE 3

ROTATED FACTOR MATRIX FOR THE VALIDATION SAMPLE

Item and content	Factor				Communality
	I	II	III	IV	
141 ... sits more comfortably	−.37	.20	—	−.30	.30
155 ... feel suffocated	—	—	−.64	.26	.49
174 ... feel constraining	.47	−.21	−.62	.21	.69
242 ... feel confined	.31	—	−.77	.21	.77
125 ... often worry ... accident	—	.70	—	—	.54
151 ... can not help worrying	—	.46	−.41	—	.41
238 ... always worry ... family is driving	—	.45	−.28	—	.33
270 Before driving ... often worry	—	.62	—	—	.45
231 ... prevent casualities	—	—	—	−.68	.49
245 ... effective ... against brain damage	—	—	—	−.57	.35
261 It has been proven ... escape injuries	—	—	—	−.57	.39
265 ... by preventing being thrust out	—	—	—	−.50	.27
158 ... gets caught in the door	.22	—	−.33	—	.16
163 ... rather simple to put on	−.90	—	—	—	.85
227 ... trouble putting on	.76	—	−.30	—	.69
257 ... trouble to adjust	.64	—	—	—	.46

were notified by mail and later on they were visited in their homes by interviewers.

Results

Beliefs about Seat Belt Usage

A factor analysis of the 16 items, constituting the four factor scores that gave optimal predictions of reported usage in Study 1, was performed. The factors were extracted in the same way as in Study 1, but in this case orthogonal rotation with the rotation criterion called Kaiser's varimax yielded the solution that was most easily interpreted. The rotated matrix and the communalities are given in Table 3.

The factor pattern showed close agreement with the original pattern in Study 1 with regard to the marker variables. Only two variables deviated from the previous outcome. Item 141 had its salient loading in the Inconvenience factor instead of the Discomfort factor for which it was chosen, and for Item 158 the reverse was true. Thus, on the whole, the belief structure obtained in Study 1 could be generalized to the present sample. As in Study 1 the factor scores were obtained by summation of the scores on the four variables in each factor.

Prediction of Seat Belt Use from Beliefs about Usage

Reported use. On the basis of the factor analysis for the present sample the inclusion of Items 141 and 158 in the final instrument was regarded as questionable. A recalculation of the multiple correlations for the sample in Study 1 was made with the two items removed. The correlations were at least as high as for the original 16 items.

In Study 1 the prediction that was made on only two factors, Discomfort and Effect, was nearly as good as the optimal prediction based on the four factors. The regression equation thus obtained was used to predict reported use for the present validation sample. The correlation between predicted and actual values was .61, which was at least as high as that which resulted when the prediction was based on all the four factors.

Observed use. To investigate how well beliefs could differentiate between the users and the nonusers in terms of observed usage, discriminatory analyses were performed. The point-biserial correlation between the optimally weighted sum of the four factors (14 items) and use (dichotomized) was .39. When only the two factors Discomfort and Effect were used the point-biserial correlation was

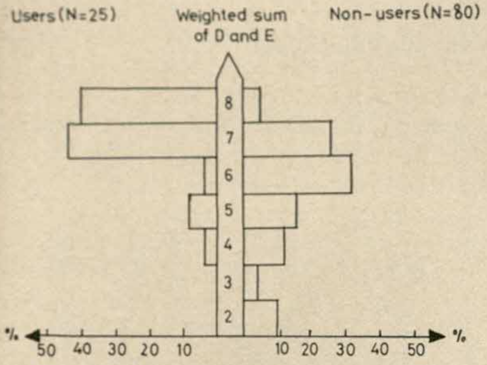


FIGURE 1. The power of the discomfort (D) and effect (E) combination to discriminate between users and nonusers.

practically the same, or .38. The result for the two factors is illustrated in Figure 1.

The difference between the two groups is clear, although the overlap between the distributions is considerable. The shortcomings of the behavior measures, which have been pointed to, and the small size of the sample of users should be kept in mind when evaluating the outcome. Variations in especially such situational factors as driving environment and distance could be important, even though the situation "driving to work" was assumed to apply to all drivers.

The weights for the optimal combination of Discomfort and Effect for the prediction of actual use were similar to those obtained for the prediction of reported use. Hence, a high correlation ($r = .83$) between the two predicted values was obtained.

DISCUSSION

The organization of beliefs with respect to seat belts and belt usage deviated rather little from what had been expected and showed a considerable stability over individuals. Similar patterns were obtained for groups of car owners differing widely in reported belt usage. Also, the belief pattern of the validation sample was almost identical to the original pattern in spite of the differences in procedures for subject selection and data collection.

The model of attitude as a sum of products of beliefs and evaluative aspects of those beliefs appeared to be useful. An independent measure of attitude (A_o) could be predicted almost as well as the reliabilities of the measures would permit from conceptions about the discomfort of wearing seat belts and their injury preventive effects.

The multiple correlation between reported usage and the combination of the Discomfort and Effect factors was almost as high as the correlation between A_o and reported usage. A small improvement of the predictions could be made by adding information concerning experience of inconvenience in handling the belts and of worry about getting injuried in a traffic accident.

For the validation sample in Study 2 the correlation between reported and predicted use, based on the combination of the Discomfort and Effect factors, was .61. It must be kept in mind that it was not a random sample, but as it showed the same distribution of background variables as the sample in Study 1, and the reported level of belt usage barely exceeded that of the Study 1 sample, the coefficient can be taken as further support for the predictive validity of the weighted sum of these two factors. The value of the coefficient is considerably higher than values that have generally been reported for the correspondence between attitudes and belt usage. Obviously, some common variance could be due to methods similarity (see Campbell & Fiske, 1959), since self-reports were used in all cases, although they were not identical in form. The Discomfort–Effect combination also had some power to discriminate between users and nonusers observed in the situation "driving to work." The usefulness of the obtained relationships for the construction of a model of seat belt use was thus further demonstrated.

An illustration of the relationships and variables that seem most important on the basis of the present studies is given in Figure 2.

On the basis of the intercorrelations given in Figure 2(a), a great number of causal models could be suggested, even if reciprocal causation and the operation of additional variables were not allowed for. The model suggested in Figure 2(b) is in line with the rationale behind many attempts to influence public behavior in various areas, a rationale which rarely has been subjected to empirical test. In this model, the latent variable, "attitude to seat belt use," is regarded as an intervening variable mediating between beliefs and actual usage, the variance being explained by variation in conceptions about the discomfort experienced when wearing seat belts and about the effects of belt usage in a car accident. The attitude, then, is assumed to be a determinant of the hypothetical variable, "disposition for belt use" (see Fhanér & Hane, 1973b).

Whether a person in a given situation will put on his belt or not will be dependent upon the relation between his value in the "disposition for belt use" variable and the characteristics of the situation. A situation can be described in terms of "difficulty," that is, the proportion of all drivers using the belt in that situation (Fhanér & Hane, 1973b). In this sense, city driving is a more difficult situation with respect to belt usage than is highway driving, since level of belt use is lower in the city.

By experimental manipulation of the Discomfort and Effect factors the value of the model proposed in this study could be assessed. It would, however, be realistic to recognize the possibility of additional determinants of the variable "disposition for belt use," even if the nonperfect correlations in part are due to measurement shortcomings. A consequence of a research strategy focusing on correlations as they exist in nature is, of course, that important causal factors can be disregarded due to low correlations. It is possible, for instance, that worry about getting involved in an accident is a determinant of "disposition for belt use." If the users reduce their worry by putting on the belt, feeling confident that the belt will reduce the negative consequences of an accident, then a zero correlation could be obtained, in spite of an existing causal relationship. It is even more reasonable, perhaps, to assume the existence of feedback loops in the proposed model. Behavior change, induced by means other than by changing attitude, for example, by law enforcement, or monetary rewards, could thus be assumed to influence the Discomfort and Effect factors.

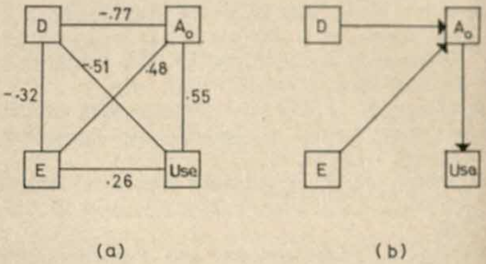


FIGURE 2. Relations between discomfort (D), effect (E), attitude (A_o), and reported use (Use). (a) Empirical relations obtained in Study 1. (b) Suggested causal links.

In summary, the relations between beliefs, attitude, and usage obtained in the present studies were substantial and definitively large enough to merit further investigation into their nature. The traditional paradigm behind traffic safety campaigns would thereby be given an empirical evaluation.

REFERENCES

Allport, G. W. Attitudes. In M. Fishbein (Ed.), *Readings in attitude theory and measurement.* New York: Wiley, 1967. (Originally published: In C. Murchison (Ed.), *Handbook of social psychology.* Worcester, Mass.: Clark University Press, 1935.)

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 1959, **56,** 81–105.

Fhanér, G., & Hane, M. *Seat belts: Contextual factors and bias of reported use.* (Report No. 23) Solna, Sweden: The National Swedish Road Safety Office, Driver license and Research Departments, 1972.

Fhanér, G., & Hane, M. Seat belts: Factors influencing their use. A literature survey. *Accident Analysis & Prevention,* 1973, **5,** 27–43. (a)

Fhanér, G., & Hane, M. Seat belts: The importance of situational factors. *Accident Analysis & Prevention,* 1973, **5,** 267–285. (b)

Fishbein, M. Attitude and the prediction of behavior. In M. Fishbein (Ed.), *Readings in attitude theory and measurement.* New York: Wiley, 1967. (a)

Fishbein, M. A consideration of beliefs and their role in attitude measurement. In M. Fishbein (Ed.), *Readings in attitude theory and measurement.* New York: Wiley, 1967. (b)

Fishbein, M., & Raven, B. H. The AB Scales: An operational definition of belief and attitude. In M. Fishbein (Ed.), *Readings in attitude theory and measurement.* New York: Wiley, 1967.

Jennrich, R. I., & Sampson, P. F. Rotation for simple loadings. *Psychometrika,* 1966, **31,** 313–323.

McGuire, W. J. The nature of attitude and attitude change. In G. Lindzey & E. Aronson (Eds.), *Handbook of Social Psychology*, Vol. 3 (2nd ed.). Reading, Mass.: Addison-Wesley, 1969.

Rosenberg, M. J. Cognitive structure and attitudinal effect. *Journal of Abnormal and Social Psychology*, 1956, **53**, 367–372.

Rosenberg, M. J. A structural theory of attitude dynamics. *Public Opinion Quarterly*, 1960, **24**, 319–340.

Sjöberg, L. *Factor analysis and multidimensional scaling, I–IV*. Göteborg: University of Göteborg, Department of Psychology, 1971–1972.

Statistiska Centralbyrån. Folk- och bostadsräkningen, 1970. (The National Central Bureau of Statistics. The 1970 census.) Stockholm: Statistiska Centralbyrån.

Thurstone, L. L. The measurement of social attitudes. *Journal of Abnormal and Social Psychology*, 1931, **26**, 249–269.

Wicker, A. W. Attitudes versus actions: The relationship of verbal and overt behavioral responses to attitude objects. *Journal of Social Issues*, 1969, **25**(4), 41–78.

(Received July 16, 19[illegible])

Journal of Applied Psychology
1974, Vol. 59, No. 4, 483–489

EFFECTS OF CAFFEINE ON ALERTNESS IN SIMULATED AUTOMOBILE DRIVING[1]

EDMUND G. REGINA AND GENE M. SMITH[2]
Massachusetts General Hospital, Boston, Massachusetts

CHARLES G. KEIPER[3] AND ROBERT K. McKELVEY
Injury Control Research Laboratory, U.S. Department of Health, Education, and Welfare, Providence, Rhode Island

Thirty minutes after ingesting 200 milligrams of caffeine or a placebo, each of 24 male subjects drove an automobile simulator for 90 minutes. Immediately thereafter, the subject ingested a supplemental dose of the medication taken initially (200 milligrams of caffeine or placebo), and then drove for another 90 minutes. The simulator provided a comprehensive and coherent set of stimulus inputs which produced a degree of realism not usually found in laboratory studies. Both the initial and the supplemental doses of caffeine significantly enhanced performance beyond that found with placebo, on each of the four measures of alertness.

The belief that caffeine enhances mental and physical performance is widespread and deeply embedded in our folklore, but experimental support for this belief is neither extensive nor unequivocal. In a comprehensive review of experimental studies, Weiss and Laties (1962) reported that caffeine seems to enhance some aspects of performance but that the evidence regarding the effect on human performance is not conclusive. Fleishman and Bartlett (1969), who reviewed recent research on the effects of drugs on human performance, pointed out the need for experimental studies evaluating the effects of specific drugs and dosage on specific classes of human performance.

In view of the prevalence of the use of caffeine (often for the specific purpose of enhancing alertness upon awakening in the morning, when studying, or driving an automobile), scarcity of systematic information about its effects on performance is somewhat surprising. For example, in searching the literature, we encountered only three well-controlled studies of the effects of caffeine on complex performance related to operation of vehicles. Two of these studies (Hauty & Payne, 1955; Payne & Hauty, 1954) demonstrated that caffeine can enhance monitoring performance in an aviation-training device. The other study (Baker & Theologus, 1972) showed that caffeine can enhance performance on a visual monitoring task analogous to an aspect of automobile night driving.

The purpose of the present study was to evaluate the effects of caffeine on four indices of performance in an automobile-driving simulator. The indices were based on time of response to accelerations and decelerations of a lead car the subject was following and on the speed and dependability of his responses to a high-beam signal. The unique facilities available at the Department of Health, Education, and Welfare (DHEW) laboratory where this study was conducted made it possible to combine the control and precision of laboratory research with a highly realistic simulation of actual highway driving.

METHOD

Subjects

The subjects were 24 healthy males ranging in age from 21 to 26 years, whose vision was mea-

---

[1] This work was performed at the Injury Control Research Laboratory, U.S. Department of Health, Education, and Welfare, Providence, R.I., and was supported in part by United States Public Health Service Grant GM-15904 and in part by a grant from the Bristol-Myers Company.

[2] The authors are grateful to William G. Cochran and Allen Donner for their guidance in statistical matters, and to Marion C. Oviatt and other staff members at the Department of Health, Education, and Welfare Injury Control Research Laboratory for their help in this study.

Requests for reprints should be sent to Gene M. Smith, Massachusetts General Hospital, Boston, Massachusetts 02114.

[3] Now at the Department of Transportation, Alcohol Countermeasures Program, Washington, D.C.

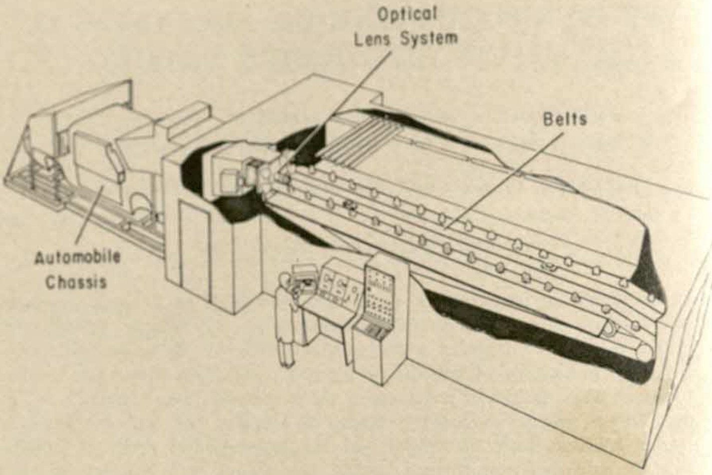


FIGURE 1. Automobile driving simulator.

sured as 20/30 or better. All were experienced drivers, and none drank more than five cups of coffee per average day. The subject was told that the experiment concerned the effects on driving performance of a moderate dose of a commonly used, nonprescription agent. After completing a 45-minute practice drive in the simulator, the subject was invited to participate in the study on each of two experimental days. Subjects were asked to get a normal night's sleep before each experimental day, to avoid use of mood-altering drugs (including caffeine) the evening before and the day of each experimental session, and not to eat for a period of two hours prior to each session.

Apparatus and Experimental Tasks

The automobile-driving simulator (see Figure 1), which was constructed around the chassis of a 1965 Rambler with automatic transmission, presented a realistic situation in which the driver performed the operations required in actual automobile driving. The driver received sensory input simulating that which would be experienced by an individual driving at twilight along a straight, continuous, flat road in a rural environment. The driver interacted with a lead car by accelerating, decelerating, and braking as he believed the situation required. In addition, he responded to aperiodically appearing high-beam signals, as described below, and carried out these activities while continually steering the Rambler to maintain correct lateral position on the roadway.

Softly lighted instruments and hidden light sources provided low-level interior lighting. Visual displays mounted on the side and rear windows of the chassis provided a static simulation of a highway environment at twilight.

The chassis stood at one end of five adjacent movable belts, each of which was a 70-foot loop. The belt system simulated two driving lanes, the centerline, and the roadsides (scenery) of a rural highway. An optical lens system mounted on the front of the car made the scale model of the roadway appear life-size to the driver. The resulting visual display simulated approximately 750 feet of actual highway distance.

Simulation of forward motion was accomplished through the rotation of the centerline and roadside belts toward the driver. The rotation of these belts, controlled by the operation of the accelerator and the brake of the car, produced a visual simulation of the "approach" of road and scenery that would actually be seen in a moving car; that is, operation of the Rambler's accelerator and brake produced appropriate changes in the rate of approach of the road and scenery. Coordination of the rotation of belts with varying motor noise, wind effects, and speedometer readings provided additional sensory inputs that simulated real driving.

A plastic model of an automobile was mounted on the right-lane belt and, by virtue of the optical system, this model simulated a life-size lead car with which the driver interacted. The simulated speed of the lead car (controlled by a program recorded on magnetic tape) and the "speed" of the Rambler (controlled by the driver's operation of the accelerator and brake) determined the apparent distance between the lead car and the Rambler. In the resulting visual display, the driver could overtake the lead car, maintain a constant distance behind it, or drop back from it. The perceived closing and separation rates between the two cars simulated those experienced in actual driving at corresponding speed differences. During each experimental drive, the lead car traveled at simulated speeds ranging from 45 to 65 miles per hour, making five accelerations and five decelerations aperiodically during each 30-minute period. The subject's task was to follow the lead car at a distance of one car length for each 10 miles per hour of speed and

TABLE 1

RESPONSE TIME TO LEAD CAR DECELERATIONS

Dose	Drug condition	$\bar{X}$ (seconds)		SD[a]	t	% improvement[b]
		Logarithm	Antilogarithm			
Initial	Caffeine	.616	4.14	.114	2.32*	12
	Placebo	.670	4.68			
Supplemental	Caffeine	.614	4.11	.094	4.22**	17
	Placebo	.695	4.96			

[a] The standard deviation of caffeine–placebo differences (in logarithms) is based on analysis of variance for a cross-over design (Cochran & Cox, 1950, p. 115) with variance due to order extracted from the estimate of variability; $df = 22$.

[b] Percent improvement is reported in terms of antilogarithms of means following the formula: $\frac{C - P}{P}$ (100).

* $p < .05$.
** $p < .001$.

to correct his speed whenever he recognized an increase or a decrease in the speed of the lead car.

High-beam signals were controlled by a mechanical programmer which activated the high-beam indicator light on the Rambler's dashboard. Depressing the dimmer switch on the car's floorboard extinguished the light. If not extinguished by the driver, the light was automatically turned off after 3.2 seconds. The aperiodic appearance of the high-beam light was explained to the subject by stating that a simulated defect in his headlights would occasionally cause them to switch from low beam to high beam, automatically and noiselessly. He was instructed to dim his lights as rapidly as possible, whenever the high-beam indicator light appeared. During each 30 minutes of driving, eight high-beam signals were presented.

Design

In this comparison of orally administered caffeine and placebo, each subject served as his own control in a double-blind, counterbalanced study in which 12 subjects received caffeine first, and 12 subjects received placebo first.

Drugs were administered in the form of white peppermint-flavored tablets taken with water. Approximately 30 minutes after the initial dose of medication (tablets containing either 200 milligrams of anhydrous caffeine or placebo), the subject began an uninterrupted drive of 90 minutes. After this, there was a break of approximately 10 minutes, during which the subject left the car, took a supplemental medication (repeat of the initial dose), and was instructed to void. Thereupon the subject returned to the car and began a second 90-minute drive.

Because the dose-supplementation procedure represents an unconventional experimental design, its adoption for this study merits comment. The unusual combination of realism and laboratory precision provided by the driving simulator made it especially attractive for our study, but the considerable demand for its use limited its availability for any one particular study. Initially, we had intended to select an experimental design which permitted a test of two or more dose levels, in order to obtain information about dose-response relationships and also to reduce the risk of failing to find real but subtle caffeine effects. However, the restricted availability of the simulator forced us to limit our study to approximately 50 experimental trials. Thus, with a two-dose design (e.g., 200 milligrams, 400 milligrams, and placebo), we could have studied only 16 subjects; while with a one-dose design (200 milligrams of caffeine and placebo), we could have studied 24. Although we chose the latter alternative, we did so realizing that our experiment might be too insensitive to reveal a statistically significant effect with the 200-milligram dosage. For that reason, we adopted a dose-supplementation procedure which made it possible to study a somewhat larger dose (as well as the 200-milligram dose of caffeine) without increasing the number of experimental sessions required. By administering 200 milligrams of caffeine as a supplemental dose two hours after the first dose, it was possible to obtain information about the effects of 200 milligrams of caffeine during the first 90-minute driving period and to obtain information about a higher level during the final 90-minute driving period. Although the effective level of caffeine during the final 90-minute drive cannot be specified in this study, data reported by Axelrod and Reichenthal (1953) concerning the fate of caffeine in man, made it reasonable to assume that blood plasma levels of caffeine during the final 90-minute drive would be higher than those during the first 90-minute drive.

RESULTS

Performance was evaluated according to four indices of alertness: (*a*) response time

TABLE 2

RESPONSE TIME TO LEAD CAR ACCELERATIONS

Dose	Drug condition	$\bar{X}$ (seconds)		SD[a]	t	% Improvement[b]
		Logarithm	Antilogarithm			
Initial	Caffeine	.664	4.61	.115	2.56*	13
	Placebo	.724	5.30			
Supplemental	Caffeine	.615	4.12	.124	6.56**	32
	Placebo	.781	6.04			

[a] Standard deviation of caffeine-placebo differences (in logarithms) is based on analysis of variance for a cross-over design (Cochran & Cox, 1950, p. 115) with variance due to order extracted from the estimate of variability; $df = 22$.

[b] Percent improvement is reported in terms of antilogarithms of means following the formula: $\frac{C - P}{P}$ (100).

* $p < .05$.
** $p < .001$.

to lead car decelerations; (*b*) response time to lead car accelerations; (*c*) response time to presentations of the high-beam signal; and (*d*) the number of missed high-beam signals. The effects of caffeine were assessed after an initial 200-milligram dose of caffeine (the first 90-minute drive) and after a supplemental 200-milligram dose (the second 90-minute drive). In all analyses, response times were transformed to logarithms.

Lead car decelerations. Response time to a lead car deceleration was measured from the start of the deceleration to the first subsequent release of the Rambler's accelerator by the subject. For both the initial and the supplemental medication treatments, responses to lead car decelerations were significantly faster after subjects had received caffeine than after placebo (see Table 1). The improvement in response time was 12% after the initial dose of 200 milligrams of caffeine and 17% after the supplemental dose of 200 milligrams of caffeine.

Lead car accelerations. Response time to a lead car acceleration was measured from the start of the acceleration to the first subsequent depression of the Rambler's accelerator by the subject. For both drug levels, responses to lead car accelerations were significantly faster after caffeine than after placebo (see Table 2). The improvement in response time was 13% after the initial caffeine dose and 32% after the supplemental dose.

High-beam signals. Responses to a high-beam signal, as measured by the time interval between the onset of the high-beam indicator light and activation of the dimmer switch, was significantly faster after caffeine than after placebo (see Table 3). (Missed signals did not contribute to the subject's response time score.) There was an 11% improvement in response time after the initial 200-milligram dose of caffeine and a 1[illegible]% improvement after the supplemental dose.

Missed high-beam signals. Compared with the placebo condition, the subjects missed 30% fewer signals after the initial dose of caffeine and 50% fewer signals after the supplemental dose. Both differences were statistically significant (see Table 4).

Comparison of 90-minute periods. As already indicated by the percentage of improvement figures, the drug effect (caffeine–placebo difference) was somewhat greater during the second 90-minute drive than during the first. This was not surprising in view of the dose-supplementation feature of the experimental design. Although each of the four indices showed a trend toward a larger effect during the second than during the first 90-minute drive, the difference was statistically significant with only one of the four; namely, response time to lead car accelerations ($F = 10.72$, $df = 1/115$, $p < .005$).

Analyses for 30-minute comparisons. As shown in Table 5, none of the four caffeine–

TABLE 3

RESPONSE TIME TO HIGH-BEAM SIGNALS

Dose	Drug condition	$\bar{X}$ (seconds)		SD[a]	t	% improvement[b]
		Logarithm	Antilogarithm			
Initial	Caffeine	.162	1.45	.072	3.40*	11
	Placebo	.212	1.63			
Supplemental	Caffeine	.158	1.44	.085	3.80**	14
	Placebo	.224	1.68			

[a] Standard deviation of caffeine-placebo differences (in logarithms) is based on analysis of variance for a cross-over design (Cochran & Cox, 1950, p. 115) with variance due to order extracted from the estimate of variability; $df = 22$.

[b] Percent improvement is reported in terms of antilogarithms of means following the formula: $\frac{C-P}{P}(100)$.

* $p < .01$.
** $p < .001$.

placebo comparisons for the first 30 minutes of the first [illegible]-minute drive was statistically significant. Of the 20 remaining 30-minute comparisons, [illegible] were in the direction of caffeine enhanced alertness, and 14 were statistically significant. These analyses indicate (*a*) the drug effect was definite enough to yield significant caffeine–placebo differences with rather small time units, that is, 30 minutes, and (*b*) the statistical significance of the main results reported in Tables 1–4 rests on drug effects appearing throughout the experiment, after the first 30-minute interval of the first drive.

The breakdown of the data into 30-minute intervals also permits inspection of changes in the drug effect (caffeine–placebo difference) over time. Table 6 summarizes the analysis of variance performed on each of the four indices to test these changes. With two of the four indices (response times to lead car accelerations and to high-beam signals) significant differences in the drug effect among the six intervals were found. This was due mainly to a progressive increase in drug effect over time (perhaps caused by progressive fatigue, or increased absorption of caffeine, or both) and partly to changes in drug effect over the break. Comparing the drug effects (caffeine–placebo difference) obtained in the 30-minute intervals immediately preceding and immediately following the break showed results which were not uniform from index to index (see Table 5). The drug effect increased after the break with two indices (response time to decelerations and missed high-beam signals), remained about the same with one (response time to accelerations), and decreased with the fourth (response time to high-beam signals). The last mentioned index was the only one for which change across the break was statistically significant ($F = 4.28$, $df = 1/115$, $p < .05$).

In general, the analyses of separate 30-minute intervals failed to reveal simple time-related drug effects which were common to the four indices. Moreover, the dose-supplementation procedure (which was adopted to optimize the chances for finding a significant main effect with caffeine) precludes straightforward interpretation of any time-related

TABLE 4

MISSED HIGH-BEAM SIGNALS

Dose	Drug condition	$\bar{X}$ number missed	SD[a]	t	% improvement[b]
Initial	Caffeine	3.25	3.19	2.12*	30
	Placebo	4.63			
Supplemental	Caffeine	2.75	3.71	3.63**	50
	Placebo	5.50			

[a] Standard deviation of caffeine–placebo differences is based on analysis of variance for a cross-over design (Cochran & Cox, 1950, p. 115) with variance due to order extracted from the estimate of variability; $df = 22$.

[b] Percent improvement is reported following the formula: $\frac{C-P}{P}(100)$.

* $p < .05$.
** $p < .01$.

TABLE 5

CAFFEINE AND PLACEBO MEANS FOR EACH OF SIX 30-MINUTE COMPARISONS OF FOUR INDICES OF ALERTNESS

Drug	Driving intervals after initial dose[a]			Driving intervals after supplemental dose[a]		
	Minutes					
	0–30	30–60	60–90	0–30	30–60	60–90
	Response time to lead car decelerations					
Caffeine	.587	.600	.642***	.596	.604	.626
Placebo	.639	.663	.684	.696	.666	.684
	Response time to lead car accelerations					
Caffeine	.671	.642*	.649*	.562**	.625**	.628***
Placebo	.662	.725	.753	.702	.784	.801
	Response to high-beam signals					
Caffeine	.145	.177**	.149**	.188	.142**	.129***
Placebo	.155	.237	.228	.198	.235	.220
	Missed high-beam signals					
Caffeine	.83	1.04*	1.38	1.12*	.67**	.96**
Placebo	1.25	1.71	1.67	1.88	1.88	1.75

[a] Response time means are in logarithms. The error term used for each t test was obtained as indicated in the footnote to Table 1; $df = 22$.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

changes in drug effect that might be suggested by the results reported in Table 5. This study clearly demonstrates that caffeine, when compared with placebo, did produce improved performance on all indices, but inferences concerning more complex caffeine effects (such as interactions among effects of drug, time, and performance index) are, we believe, unwarranted.

DISCUSSION

The results of this study are consistent with reasonable expectations and with empirical findings of three earlier studies (Baker & Theologus, 1972; Hauty & Payne, 1955; Payne & Hauty, 1954) concerning the effects of caffeine on tasks involved in vehicle operation. Perhaps the most important aspect of the present study is that the effects of caffeine were evaluated in a situation which provided the advantages of laboratory control without undue sacrifice of realism. Laboratory findings are often limited by their questionable transfer to events outside the laboratory setting. On the other hand, attempts to study phenomena in natural settings often sacrifice manipulative control and precision of measurement.

The present study was conducted in an environment that provided a comprehensive and coherent set of stimulus inputs which produced a degree of realism not usually found in laboratory studies. The subject sat inside a full-sized automobile and received visual, auditory, and tactile inputs associated with being so situated. This was augmented by a simulation of the exterior environment involving such things as motor sounds, wind effects, and a visual field that included onrushing scenery and a continuously moving

TABLE 6

ANALYSES OF VARIANCE FOR THE CAFFEINE–PLACEBO DIFFERENCE OVER SIX INTERVALS

Source	*df*	*MS*ᵃ	*F*
Response time to lead car decelerations			
Intervals (I)	5	.010	.48
Subjects (S)	23	.037	1.76
Error (I × S)	115	.021	
Response time to lead car accelerations			
I	5	.107	3.34**
S	23	.050	1.56
I × S	115	.032	
Response time to high-beam signals			
I	5	.035	2.69*
S	23	.015	1.15
I × S	115	.013	
Missed high-beam signals			
I	5	2.48	1.23
S	23	3.86	1.91
I × S	115	2.02	

ᵃ Response time data are in logarithms.
* $p < .05$.
** $p < .01$.

highway centerline. Within this setting, the caffeine-enhanced performance was broad in nature, encompassing responses to both gradually appearing signals (changes in lead car speed) and signals which were discreet (on-off high-beam signals). Moreover, the lead car signals came from the roadway in front of the driver's car, while high-beam signals, which were relatively unobtrusive, came from the instrument panel inside the car.

The somewhat unconventional two-stage, dose-supplementation procedure was selected for this study with the expectation that caffeine-induced alertness would be easier to detect later than earlier during the three hours of driving, because of progressive increases in caffeine concentration in the blood. The observed tendency for the caffeine–placebo difference to be greater during the second than during the first driving period is consistent with that expectation. Moreover, the fact that the caffeine–placebo differences for the two 90-minute drives were consistent with each other adds reassurance to the conclusion of a dependable caffeine–placebo difference. The results for the second drive cannot be said to comprise an independent replication of those for the first, but the consistency observed does provide a kind of cross-validation, and this adds to the confidence with which we conclude that the behavioral parameters investigated in this study are dependably altered by caffeine. Similarly, the fact that 14 of the 20 thirty-minute comparisons made after a 60-minute absorption period showed statistically significant superiority of caffeine over placebo also adds to the confidence of that conclusion.

REFERENCES

Axelrod, J., & Reichenthal, J. The fate of caffeine in man and a method for its estimation in biological material. *Journal of Pharmacology and Experimental Therapeutics*, 1953, **107**, 519–523.

Baker, W. J., & Theologus, G. C. Effects of caffeine on visual monitoring. *Journal of Applied Psychology*, 1972, **56**, 422–427.

Cochran, W. G., & Cox, G. M. *Experimental designs*. New York: Wiley, 1950.

Fleishman, E. A., & Bartlett, C. S. Human abilities. *Annual Review of Psychology*, 1969, **20**, 349–380.

Hauty, G. T., & Payne, R. B. Mitigation of work decrement. *Journal of Experimental Psychology*, 1955, **49**, 60–67.

Payne, R. B., & Hauty, G. T. The effects of experimentally induced attitudes upon work proficiency. *Journal of Experimental Psychology*, 1954, **47**, 267–273.

Weiss, B., & Laties, V. G. Enhancement of human performance by caffeine and the amphetamines. *Pharmacological Reviews*, 1962, **14**, 1–36.

(Received July 25, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 4, 490-496

TIME DELAY AND SIMILARITY EFFECTS IN FACIAL RECOGNITION[1]

KENNETH R. LAUGHERY[2]
University of Houston

PAUL K. FESSLER, DAVID R. LENOROVITZ, DARRYL A. YOBLICK
State University of New York at Buffalo

Four experiments investigated the ability of undergraduate students to recognize a target person (whom they had previously seen in slides or a film) in a test series of pictures. The first study, which manipulated time delay between seeing the target and starting the test series (from four minutes to one week), demonstrated that delay had no effect on recognition performance. The other three studies manipulated similarity between the target person and the decoy pictures preceding the target in the search series. The greater the target-decoy similarity, the poorer the recognition performance. A significant target-position-similarity interaction indicated the presence of a larger performance decrement due to position when similarity level was high.

In an earlier series of experiments Laughery, Alexander, and Lane (1971) explored the effects of several task variables on facial recognition. The paradigm of those studies resembled the situation in which a witness who has seen a criminal tries to identify that person's picture from a set of alternatives. The results indicated that one variable in particular, target position, had a strong effect on recognition performance. Recognition performance decreased as the number of decoy pictures preceding the target increased. As Laughery et al. (1971) pointed out, the effects of target position could be accounted for either on the basis of interference due to the number of decoy faces intervening between the original target exposure and the appearance of the target in the search series, or on the basis of the greater amount of time elapsing when the target appeared in a later position.

The issue of which of these two explanations more appropriately accounts for the results of the position effect is, of course, a long-standing problem in psychology. Studies that have manipulated the amount of intervening material between training and testing in retention studies have traditionally confounded interference and time lag (decay) factors. While this issue does not lend itself to any clearcut experimental resolution, there are procedures which can help to gain some insight into the relative importance of the two factors. Furthermore, it should be noted that the relative weighting of these factors has important implications for the design of criminal identification systems.

Most current criminal identification files are organized strictly on the basis of sex and race. One development that might take place involves designing a personal appearance coding system that would permit faces to be classified on the basis of other characteristics (color of hair, bushiness of eyebrows, etc.). The size of the search set could then be decreased by additional descriptors of the criminal. Note, however, that such a procedure would result in a search set where the decoys are more similar to the target and presumably would produce a greater amount of interference. If interference is the major contributing factor, the additional interference caused by the more similar decoys might off-

---

[1] This research was supported in part by Research Grant No. MH-11595 from the National Institute of Mental Health, United States Public Health Service, and by Research Grant No. NI 70-065-PG-15 from the National Institute of Law Enforcement and Criminal Justice, United States Department of Justice.

[2] The assistance of David Coe in setting up laboratory equipment and Kenneth R. Laughery, Jr. in analyzing data is gratefully acknowledged.

Requests for reprints should be sent to Kenneth R. Laughery, Department of Psychology, University of Houston, Houston, Texas 77004.

set the value of the smaller search set. If time delay, on the other hand, were the more potent factor, then emphasis would be placed upon getting the witness to the identification task quickly, and the delimiting of the search set could proceed with less concern for the increased interference due to greater similarity.

This article presents the results of four experiments, one dealing with time delay and three concerning facial similarity in the identification task. The different similarity studies employed different definitions of similarity.

EXPERIMENT 1

In this experiment the time delay between the initial exposure to the target and the start of the search series was manipulated over six values ranging from four minutes to one week.

Method

Subjects. The subjects were 292 undergraduate students enrolled in an introductory psychology course at the State University of New York at Buffalo. Participation in the study partially fulfilled a course requirement.

Design. The task in this experiment was basically the same as in those reported by Laughery et al. (1971). The subjects were first exposed to a target person, viewing either four sequentially presented slides of the target person in different candid positions or viewing a one-minute movie film clip[3] of a scene in which the target person played a significant role. The subject's task was to indicate, using a 6-point scale (definitely not, probably not, possibly not, possibly yes, probably yes or definitely yes), whether each slide in a subsequent, serially presented test series was or was not the target. The target's picture appeared only once in the search series.

In the four candid position slides, the target was shown in positions ranging from left side to right side, full length, and bust views. The candid positions were obtained by taking a large number of pictures of the target, and selecting those that seemed least posed. One target was shown with glasses in both the four candid position slides and in the test series; while a second target appeared without glasses. The slides were projected so as to be approximately life-size on the screen.

The film depicted a check-cashing episode at a department store. It was prepared by WCNY television at Syracuse, New York. Five characters appeared in the film: three female customers, a male customer, and a male clerk. The male customer and the clerk, the principal characters, were the targets in the study. In the film, the clerk waited on one of the female customers, and then cashed a check for the male customer while the remaining two customers waited impatiently. During the check-cashing process, both the clerk and the male customer were seen in various positions including front and profile. Both moved about and spoke.

The test series consisted of 149 decoys and the target, all appearing in front bust views. The decoys, all white males within an age range of 17 to 23 years, were randomly selected from a picture (slide) library of 544 different faces. The physical parameters of all slides were constant (sharpness, scale, lighting, etc.). Faces of fair and darker coloring, with and without glasses, were approximately equal in number and randomly distributed throughout the series. During the search sequence, each slide was projected on the screen for four seconds with six seconds between slides—during which the subjects recorded their responses on the answer sheets.

The design of the experiment was a 6 × 2 factorial with both factors manpiulated as between-subjects variables. The first variable, time delay, was defined as the time lapse between initial exposure to the target and the start of the test series. Intervals of 4 minutes, 30 minutes, 1 hour, 4 hours, 1 day, and 1 week were used. The second variable, mode of exposure, consisted either of presenting the four casual slides or the movie scene. Two different targets were used in each of the mode conditions in order to decrease the likelihood that any obtained differences would be the result of some idiosyncratic characteristic of the target. Thus, the two variables combined with the two targets in each mode condition produced a total of 24 experimental conditions. A separate group of subjects was used in each condition, with the number of subjects per condition ranging from 9 to 14.

A point to be noted is that the two targets in the casual slides condition were not the same as the two in the movie condition. It is realized, of course, that the mode of exposure variable is thus confounded with targets. Stimulus materials were simply not available for equating the targets in the two exposure conditions. The reason for proceeding in this manner was that the movie was available, and we were interested in exploring possible mode of exposure effects. Also, a meaningful interaction between time delay and mode of exposure might suggest interesting ideas for future work. In any case, the mode of exposure variable should be considered an exploratory effort; the thrust of the study was the time delay variable.

The subjects in each of the 24 experimental conditions were run as a group, with groups of subjects randomly assigned to the various conditions. The experiment was carried out in a normal university classroom with the screen located in the front center of the room. The room was darkened to insure good vision of the slide or movie, but with sufficient light to read and mark the answer sheet.

---

[3] Further details of the apparatus and procedures employed may be obtained from the author.

The instructions were presented in two parts. The first part made it clear that the subjects would later be looking for a picture of a person whom they were about to see (in the slides or the movie). In the movie condition the subjects were told in advance which of the five characters would be the target. Following the presentation of the target, the subjects were given the second part of the instructions. This part included details about how they were to respond and a statement that the target might appear in the test series several times, only once, or not at all. In fact, the target picture appeared just once, in Position 104. Although the subjects viewed only 150 slides in the test series, the answer sheets were numbered to 200 in order to eliminate any bias that might result from anticipating the end of the series.

The critical variable in this experiment, the time delay, was introduced between the presentation of the target and the second part of the instructions. In the 4-minute condition there was no delay, and the second part of the instructions was presented immediately (it was these instructions that filled the 4-minute interval). In the 30-minute and 1-hour delay conditions the delay period was partially filled by running the subjects in a separate memory span for letters task. In the 4-hour, 1-day, and 1-week conditions, the subjects were run in the span task and then excused. They returned, of course, at the appropriate later time.

It was decided that if any subject knew the particular target person to whom he was exposed, he would be excused from the experiment. No subjects indicated knowing any of the target persons. The subjects were asked to indicate on their answer sheet if they knew any of the decoys. There was a negligible number of responses indicating any subject knew a decoy face.

Results

The responses to the target picture in the test series are referred to as the hit-miss (H-M) scores. Similarly, the responses to the 149 decoys are referred to as false alarm-correct rejection (FA-CR) scores. For a given subject the H-M score could be a single value from 1–6, and the FA-CR score could range from 149 to 894 (the sum of all other response scores). A higher H-M score and a lower FA-CR score reflect better performance.

Two analyses were carried out on the results. The first analysis was based upon each subject's response to the target picture, the H-M score. The mean H-M score for each of the 12 experimental conditions, collapsed across targets, was calculated. An analysis of variance was carried out on this data to determine the effects of the delay and mode variables. The presentation mode had a significant effect ($F = 50.23$, $df = 1/204$, $p < .001$) with performance higher in the condition where the exposure to the target consisted of viewing four casual photographs. The overall means for the casual pictures and movie scene were 4.4 and 2.[illegible] respectively. A significant delay by mode interaction ($F = 2.93$, $df = 5/204$, $p < .05$) was probably due to the fact that performance was especially poor in the 30-minute delay, movie-mode condition. (1.8 versus 4.6 in 30-minute delay, slide condition)

The second analysis was based on the FA-CR scores. The mean score per decoy (the total FA-CR score divided by 149) for each of the 12 experimental conditions was computed. The mode of presentation effect was again significant ($F =$ [illegible], $df = 1/204$, $p < .005$) with performance superior (lower FA-CR scores) in the four casual slides condition. The means for the slide and film modes were 1.19 and 1.32, respectively. Although the delay effect was not significant, the delay by mode interaction did reach significance ($F = 4.22$, $df = 5/204$, $p < .005$). This interaction is probably due to the fact that in the 4-minute delay condition there is a reversal of the mode effect; in other words, with this delay the film mode led to a better FA-CR score (1.3 film versus 1.5 slides).

Discussion

The focus of interest in this experiment was on the delay variable. Somewhat surprisingly, the results show quite clearly that the degree of time lapse between exposure to the target and the beginning of the search through the test series had no meaningful effect on performance. The finding is surprising, since it was our expectation that the longer delay would lead to lower recognition performance. Clearly, the vast literature on human *verbal* memory indicates such an effect in recalling and recognizing material. Apparently, memory for faces in this type of task does not deteriorate simply with the passage of time, at least not over a period of one week.

Given the nature of the delay by mode interaction with the H-M and FA-CR criteria, it is difficult to draw any firm conclusions about the factors underlying these effects. In the analysis based upon the FA-CR scores it may be that the relatively short 4-minute delay condition led to a differential amount of confusion in the two mode conditions. While such confusion would be reflected in the FA-CR scores, it is anything but clear as to why such differential confusion should exist.

The effects of the presentation mode variable were clear and consistent. The fact that performance is better on the recognition task after slides than after the film may seem surprising at first. It might be assumed, for example, that the dynamic presentation in the film provides more information which in turn leads to better recognition performance. The fact that the results are just the reverse, however, may mean that the two-dimensional, static, casual slides are simply more like the two-dimensional, static slides in the search sequence, consequently resulting in a better performance. Another possible explanation is that the movie involved distracting "drama" as well as nontarget faces. A third possibility, of course, is that the two targets in the slides condition were in some way more recognizable than the two targets in the movie condition. A careful analysis of the targets, however, failed to reveal any hints as to what such differences were.

The results of Experiment 1 indicate that the target-position effect reported by Laughery et al. (1971) was not due primarily to the differential passage of time. Instead, it seems reasonable to speculate that the more important factor is the interference caused by the intervening faces.

Experiment 2

The purpose of the similarity studies was to examine interference by varying the similarity between the target person and the decoys that appeared in the series before the target person. A major problem in carrying out an experiment of this type concerns the definition of similarity. Three separate definitions of similarity were employed. The first was based upon results of the time delay study reported previously. For each of the 149 decoys a mean (across subjects) FA-CR score was available that reflected the extent to which each decoy was mistaken for each target. Hence, these results could be used to define the similarity between a particular decoy and target. The second definition was based upon physical characteristics of the people whose pictures are in the slide library. The third definition was based upon a direct similarity rating given by the subjects.

Experiment 2a

Experiment 2a was a similarity study employing the definition based upon FA-CR scores from the time delay study.

Method

All phases of Experiment 2 drew from the same population of subjects as Experiment 1.

In Experiment 2a, 88 subjects had essentially the same task as that described in the time delay study with one exception. In this experiment the test series consisted of 60 slides. The delay between the exposure to the target person (again, slides or movie) and the beginning of the test series was approximately four minutes, during which the second phase of the instructions was presented to the subjects. The apparatus was exactly the same as that described in the time delay study.

The design of the present experiment consisted of two variables both of which were manipulated between subjects. Presentation mode (slides versus film) was again examined. The second variable, similarity, consisted of two levels, high and low. Experiment 2 used the same targets as Experiment 1. The two levels of presentation mode, two levels of similarity, and the two targets in each of the presentation modes resulted in a total of eight experimental conditions. Eleven subjects were run in each condition.

While the materials used in this experiment were basically the same as in the earlier study, there was a difference in the selection of decoys for the test series. Using the mean FA-CR scores, the 52 most similar and the 52 least similar decoys were determined for each of the targets. In the high-similarity conditions the 52 decoys most similar to the particular target were used as the first 52 slides in the search sequence. The low-similarity condition used the 52 least similar decoys. In all conditions the order of the 52 slides selected as decoys for that particular condition was randomized. The target occurred in Position 53, followed by seven additional slides chosen at random from the picture library.

Results

Two separate analyses were carried out on the data using the H-M and FA-CR scores. In the H-M analysis, the effect of mode was highly significant ($F = 19.03$, $df = 1/84$, $p < .001$), with superior performance again in the four casual slides condition. The mean H-M scores for the casual photograph and film modes were 4.3 and 2.7. Neither the main effect of similarity nor the Mode X similarity interaction was significant. The means for the high- and low-similarity conditions were 3.3 and 3.7, respectively.

The analysis of FA-CR scores revealed a significant similarity effect ($F = 5.47$, $df = 1/84$, $p < .025$), with means of 1.4 and 1.6 for the low- and high-similarity conditions, respectively. This finding is not particularly interesting since the two similarity conditions were defined by differences in the FA-CR scores obtained in the previous experiment. The result does, however, serve as a useful check on the success of manipulating the similarity variable.

Experiment 2b

In the second similarity study the definition of similarity was based upon nine physical characteristics of the people whose pictures were in the film library. The physical characteristics of the people in the slides had been recorded at the time the pictures were taken (hair color, hair length, age, build, eye color, glasses, moustache, beard, and length and shape of sideburns). For 315 of the white males in the film library, the number of characteristics each pair has in common was counted. A similarity matrix for these 315 pictures was prepared and used to select high- and low-similarity decoys. The values in the cells of the matrices could change from zero to nine. One point should be noted about this definition of similarity. The measurements involved simple observations or estimates on the part of the people taking the pictures; consequently, this similarity definition should only be regarded as a rough estimate.

Method

One hundred and one subjects participated in this portion of Experiment 2. The apparatus, materials, and procedure were exactly the same as Experiment 2a with one exception. This exception was the mode of presenting the target in the original exposure; all targets were presented by showing four casual photographs.

Given the similarity matrix, an effort was made to select 52 high- and 52 low-similarity decoys for a number of targets. Two pictures were considered highly similar if they had eight or nine characteristics in common, while low similarity was defined as four or less characteristics in common. It may seem inconsistent that low similarity was not defined as two or less characteristics in common. The problem was that too few pairs had so few common characteristics. Since the objective was to maximize and minimize similarity in the high and low conditions, respectively, it was decided to forego any attempt to equate the low and high conditions by selecting the high group from those with six to nine common characteristics. With these definitions of high and low similarity, it proved impossible to find 52 high- and 52 low-similarity decoys for any given target person; there simply were no targets having both 52 similar and 52 dissimilar decoys. Consequently, in this study four different targets were used in each of the similarity conditions, resulting in a total of eight conditions or groups of subjects. The number of subjects per group ranged from 10 to 16.

Results

Two analyses were carried out on the data using the H-M and FA-CR scores. Scores were collapsed across the four different targets in each of the similarity conditions. The mean H-M scores for the low- and high-similarity conditions were 5.08 and 4.42, respectively. A t test showed this difference to be significant ($t = 2.05$, $df = 100$, $p < .025$).

In the analysis based upon FA-CR scores, the mean for the low-similarity condition was 1.40, and for the high condition it was 1.44. This difference was not significant.

Experiment 2c

The definition of similarity in this study was based upon a direct rating of similarity between faces obtained from a large number of subjects.

A matrix of similarity ratings was compiled for 222 faces in our picture library by obtaining similarity ratings for pairs. This matrix of ratings served as a basis for defining the relative similarity of any two members. A minimum of nine subject ratings was available for each cell (pair).

Method

One hundred and fifty-two subjects were involved in Experiment 2c. Except for a few changes, the task, apparatus, materials, and procedure were basically the same as the earlier studies. As in the second similarity study, the initial exposure to the target always consisted of seeing four casual slides. However, one difference in procedure involved the position of the target in the search series. In this experiment the target position was manipulated as a between-subjects variable and appeared either in Position 41 or 101.

Given the similarity matrix derived from the ratings, the search sequences for the high- and low-similarity conditions were formed as follows. Each of the 222 slides was considered a potential target. For each of these pictures the other 221 were ordered on the basis of their similarity to that picture. Then for each picture the mean similarity for the 100 most similar pictures was calculated. The three slides with the highest mean similarity for these 100 most similar pairs were selected as the high-similarity, and the appropriate 100 pictures were used as the decoys in the Position 101 condition. In the same way the three slides having the lowest mean similarity for the 100 least similar pairs were selected as targets for the low-similarity conditions. For the Position 41 condition a set of 40 decoys was selected from the 100 in such a way as to result in the same mean similarity as the overall set of 100. Thus, six targets were selected: three that were similar to their decoys and three that were dissimilar to their decoys.

The design was a 2 × 2 factorial with similarity (high and low) and target position (41 and 101) both manipulated as between-subjects variables. Each subject was run on two trials. To avoid obvious confounding, the target was different in the second trial. In addition, the levels of both the similarity and the position variables were changed. In other words, if a group viewed a high-similarity sequence with the target in Position 41 for the first trial, they viewed low-similarity sequence with a different target in Position 101 for the second.

Results

Since the trials variable was confounded in the design with the other variables, and since it is always possible that some sort of differential transfer effect may exist, the data from the two trials was analyzed separately. As a rather gross check on a possible trial effect, however, t tests were run on the two trials for the H-M and FA-CR scores collapsed across all other variables. While there was virtually no difference between trials on the H-M scores, the difference was significant for the FA-CR scores ($t = 3.67$, $df = 302$, $p < .001$. The FA-CR scores were lower on the second trial. A possible explanation for this difference is a criterion shift with respect to the decoys. The practice acquired during Trial 1 allows the subjects to become more familiar with the task, perhaps learning that the target is not likely to appear frequently or that the target is not to be presented with small changes in appearance. This familiarity with the task could lead to greater confidence in rejecting decoys and, thus, a lower FA-CR score.

The mean H-M scores for each of the four experimental conditions in Trials 1 and 2 were computed. An analysis of variance showed the main effect of similarity to be highly significant ($F = 27.13$, $df = 1/148$, $p < .0001$) on both trials, with better recognition of the target in the low-similarity condition. With respect to the position variable, Position 41 led to slightly better performance on Trial 1 and significantly better performance ($F = 6.76$, $df = 1/148$, $p < .01$) on Trial 2. The Position × Similarity interaction was significant in Trial 2 ($F = 5.79$, $df = 1/148$, $p < .02$) but not in Trial 1. The data underlying the interaction indicated that the difference between Position 41 and 101 was greater in the high-similarity conditions.

An analyses of variance was carried out on the FA-CR scores for each of the trials. The main effect of similarity was significant ($F = 12.51$, $df = 1/148$, $p < .001$), in Trial 1, with better performance in the low-similarity conditions. The mean FA-CR scores for the high- and low-similarity conditions were 1.2 and 1.4, respectively. All other effects were not significant.

Discussion

The results of the three similarity studies exhibit a clear and consistent pattern. In Experiments 2b and 2c the lower similarity decoys led to a significantly higher hit rate on the target picture. While the results of Experiment 2a did not produce a significant similarity effect on the H-M score, the means were in the expected direction. An appropriate conclusion can be drawn: the greater the similarity between a target and the decoys preceding it, the lower the probability that the target will be correctly identified.

The results of the FA-CR analyses are somewhat confusing. The prediction is that higher similarity should lead to higher FA-CR scores. The effect was significant in Experiment 2a and Trial 1 of Experiment 2c, but not in Experiment 2b or Trial 2 of Experiment 2c. In the latter case the explanation may be due to the criterion shift mentioned earlier, which may have produced a performance level that washed out the difference (a ceiling effect). No explanation for the lack of a significant FA-CR effect in Experiment 2b comes to mind, although it should be noted that the difference between the means, while quite modest, is in the expected direction.

The significant Position × Similarity interaction on Trial 2 of Experiment 2c is most interesting and consistent with the arguments presented in the present context. If each similar decoy is contributing more interference than each dissimilar decoy, one would expect a larger performance decrement due to position when the similarity level is high. This is precisely what happened.

GENERAL DISCUSSION

Returning to the issue of interference versus decay posed in the introductory section of this article, these studies clearly indicate that the important variable in facial recognition is the number and similarity of the faces preceding the target picture in the search series and not the amount of time elapsed. Indeed, the results of the third similarity study, 2c, show that the similarity between the decoy and target faces is more important than the number of decoy faces.

The results have definite implications for the design of criminal identification systems involving a linear search through a series of pictures. Reducing the size of the search set by eliminating alternatives that differ from the target on the basis of certain descriptors, will result in a more homogeneous, higher similarity set. This higher similarity could produce a level of interference that more than offsets the desirable effect of fewer alternatives.

One possible direction that identification systems might take is to move away from the linear search procedure. For example, an interactive search might be developed. Such a procedure would involve showing the witness a picture from the file, the witness reacting by noting differences between the target and the person in the picture, and a new alternative selected on the basis of these differences would be shown to the witness. The process would continue in this manner until an identification is made or a termination criterion is reached. This type of interactive approach based upon a "zeroing in" or "tree-like" search procedure is considerably more powerful than the linear search. The data from the experiments reported in this article indicate that the consideration of such new procedures is clearly warranted.

REFERENCE

Laughery, K. R., Alexander, J. F., & Lane A. B. Recognition of human faces: Effects of target exposure time, target position, pose position, and type of photograph. *Journal of Applied Psychology*, 1971, **55**, 477–483.

(Received July 16, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 4, 497-499

SHORT NOTES

RELATIONSHIP OF ROLE CONFLICT AND ROLE AMBIGUITY TO JOB INVOLVEMENT MEASURES

W. CLAY HAMNER[1] AND HENRY L. TOSI

Graduate School of Business, Michigan State University

Measures of role conflict, role ambiguity, and various job involvement variables were obtained for 61 high-level managers. Role conflict was negatively related to the amount of reported influence and positively related to the amount of perceived threat and anxiety; while role ambiguity was negatively correlated with job satisfaction and influence, and positively related to job threat and anxiety. It is suggested that organizational level be taken into account when studying the relationship of role stress factors with job involvement measures.

Kahn, Wolfe, Quinn, Snoek, and Rosenthal (1964) proposed a model which suggests that organizational factors (the employee's environment) affect the expectations of individuals concerning role behavior. Role theory posits that when the behaviors expected of an individual by others in the organization are inconsistent, he will be in a state of *role conflict* and will experience stress, become dissatisfied, and perform less effectively than if the expectations did not conflict. Role theory also predicts that *role ambiguity* will result from a lack of information about the role tasks. If an employee does not know what his duties are, what authority he possesses, or how he is to be evaluated, he may hesitate to make decisions and will have to rely on trial and and error learning in meeting the expectations of the organization.

The theoretical relationships regarding role conflict argued in the preceding paragraph are supported by a body of empirical literature. Kahn et al. (1964) found that high levels of role conflict were related to (*a*) low levels of job satisfaction, (*b*) low confidence in the organization, and (*c*) a high degree of job-related tension. Gross, Mason, and McEachern (1958) found a significantly negative correlation between perceived role conflict and three measures of job satisfaction. Tosi (1971) found that role conflict was positively related to job threat and anxiety and significantly related in a negative direction to satisfaction with the job.

Role ambiguity has also been examined in the same fashion. Just as with role conflict, Kahn et al. (1964) found that high levels of role ambiguity were related to (*a*) low levels of job satisfaction, (*b*) low confidence in the organization, and (*c*) a high degree of job-related tension. Cohen (1959) concluded that ambiguity of the situation and inconsistencies of directions raised the anxiety of subordinates, caused a less favorable attitude toward supervision, and lowered productivity.

When Rizzo, House, and Lirtzman (1970) examined role conflict and ambiguity they found that both tended to correlate (*a*) weakly, but positively with anxiety and propensity to leave the organization and (*b*) negatively with influence in the organization. Hence, it appears that role ambiguity and role conflict result in undesirable consequences for organization members and likely, therefore, for the organization itself.

Some contradictory findings regarding the role conflict and ambiguity variables have been reported. While Rizzo et al. (1970) found role ambiguity correlated significantly with low satisfaction, their study revealed little support for a significant relationship between job satisfaction and role conflict. This finding contrasted with Tosi's (1971) study which found just the opposite: no relationship between conflict and satisfaction.

In the light of these conflicting findings, the specific purpose of this research was to examine the relationship of role conflict and role ambiguity to various job involvement variables. It is hypothesized that, as in the Kahn et al. model, high levels of role conflict and role ambiguity are related to (*a*) low levels of job satisfaction, (*b*) low levels of job participation, and (*c*) high levels of perceived threat and anxiety. This study also examined the relationship of role conflict and role ambiguity to propensity to leave the organization. The relationships among the work involvement variables were also investigated.

---

[1] Requests for reprints should be sent to W. Clay Hamner, Graduate School of Business Administration, Eppley Center, Michigan State University, East Lansing, Michigan 48824.

METHOD

Subjects

The data were collected from 61 high-level managers participating in an executive development program from many industries in a large midwest urban region.

Instruments

A 10-item instrument developed by Rizzo et al. (1970) was used to measure subjects' perceptions of role conflict and role ambiguity. Rizzo et al. reported an estimate of internal consistency of about .80 for role ambiguity and .82 for role conflict. The Role Conflict scale emphasized the incompatibilities or incongruities arising out of perceptions or performance of role requirements, while the Role Ambiguity scale stressed the clarity of behavioral requirements.

A set of subscales was developed to measure the work involvement variables: The *Job Satisfaction* subscale used in this study was developed by Vroom (1963). The subjects responded to a 5-point scale with degrees ranging from "very much" to "very little" on the three-item scale. Vroom reported test–retest reliabilties of .75.

A *propensity to leave* measure developed by Rizzo et al. (1970) was used to ascertain the plans for and conditions under which the respondent would leave the organization, each item ranging from low to high propensity to leave.

Participation was measured by the four-item scale used by Vroom (1963). While no reliabilities were computed in this study, Vroom reported test–retest reliabilities of .63. The items measured the amount of influence the executive had on the decisions made in the organization.

The *job threat and anxiety* measure is an a priori scale developed by Tosi (1971). It was designed to measure the manager's concern about his job as it may be affected by conditions in the future over which he may have little or no control.

The variables analyzed were all derived from a single combined instrument of approximately 70 items.

RESULTS

Correlation coefficients were derived by testing the significance of the relationship between various measures of job involvement and role stress variables. Role conflict was not significantly related to either job satisfaction ($r = .13$) or the propensity to leave the organization ($r = .07$). There was, however, a trend toward a negative relationship to the amount of reported influence (participation) ($r = -.19$, $p < .10$), while it was positively related to the amount of perceived threat and anxiety ($r = .27$, $p < .05$)

Role ambiguity, on the other hand, was negatively correlated with job satisfaction ($r = -.25$, $p < .05$) and was positively correlated with job threat and anxiety ($r = .33$, $p < .01$). As with role conflict, role ambiguity was not significantly correlated with the propensity to leave the organization ($r = -.07$).

Job satisfaction tended to relate negatively to the perceived threat and the anxiety produced by the job ($r = -.19$, $p < .01$), and was positively correlated with the amount of upward influence perceived by the employee ($r = .32$, $p < .01$). While job satisfaction was not significantly related to the propensity to leave the organization ($r = -.12$), the executives who had the least influence were the ones who expressed a greater propensity to leave the organization ($r = -.23$, $p < .05$).

DISCUSSION

The results are consistent with the majority of other correlational studies that show partial support for the Kahn et al. model (1964). Role conflict showed the predicted relationships to participation, perceived threat and anxiety, while role ambiguity also correlated with job satisfaction, perceived threat and anxiety as predicted. Job satisfaction, surprisingly, did not correlate with either role conflict or the propensity to leave the organization. Perhaps this latter finding was due to the fact that the managers perceived role conflict as a "given" in an executive job, and because this conflict was expected, it did not produce dissatisfaction and a desire to leave the organization.

The executives who had the least influence in the organization were the ones who expressed the greatest desire to leave. This finding does not seem unusual in that the role of an executive generally carries with it decision-making powers. If this influence is not present, it could be expected that the individual would have a greater urge to move than he would when he has a great deal of influence in the organization decision-making process.

Since it has been determined that role stress is multidimensional, it is not surprising that the components of role conflict and role ambiguity would act in different ways under different conditions. This study supported the findings of Rizzo et al. (1970) and House and Rizzo (1972) which showed role ambiguity was related to job satisfaction, while role conflict was not. The findings of this study conflicted with the findings of Tosi (1971) and Tosi and Tosi (1970) who reported that role conflict was related to job satisfaction, while role ambiguity was not.

By taking a closer look at these studies, it appears that these results are not inconsistent. Instead, one could make the argument that the component of stress, which is related to the level of job satisfaction, is a function of organizational level. We make this hypothesis based on the nature of the samples studied. Both the loan officers (Tosi, 1971) and the teachers (Tosi & Tosi, 1970) are at the lowest level of their respective organizations. In fact, they are the ones performing the primary operating functions. Given that they know how to perform their jobs and that their jobs are well defined for them, role ambiguity is probably not a major problem for these lower order participants. They are much more likely to face conflicting role demands from others, and therefore, for these members role conflict could be expected to present the greatest problems and the greatest negative reactions.

For individuals in higher level positions, especially those in managerial roles, as is clearly the case for the subjects in this study and in the Rizzo et al. (1970) and House and Rizzo (1972) samples, their primary problems may arise from lack of clarity in their job. Since it has been argued by many that the managerial job is one of solving unstructured problems, or introducing more certainty into lower organization levels, it could be reasonably argued that the effects of higher ambiguity levels are more important determinants of satisfaction than are conflicting role pressures.

In summary, we conclude that future investigators should take into account the organizational level when studying the relationship of role stress factors with job involvement measures.

REFERENCES

Cohen, A. R. Situational structure, self-esteem, and threat-oriented reactions to power. In D. Cartwright (Ed.), *Studies in social power*. Ann Arbor: University of Michigan Press, 1959.

Gross, N., Mason, W. S., & McEachern, W. W. *Explorations in role analysis*. New York: Wiley, 1958.

House, R. J., & Rizzo, J. R. Role conflict and ambiguity as critical variables in a model of organizational behavior. *Organizational Behavior and Human Performance*, 1972, **7**, 467–505.

Kahn, R. L., Wolfe, D. M., Quinn, R. P., Snoek, J. D., & Rosenthal, R. A. *Organizational stress: Studies in role conflict and ambiguity*. New York: Wiley, 1964.

Rizzo, J. R., House, R. J., & Lirtzman, S. I. Role conflict and ambiguity in complex organizations. *Administrative Science Quarterly*, 1970, **15**, 150–163.

Tosi, H. Organizational stress as a moderator of the relationship between influence and role response. *Academy of Management Journal*, 1971, **14**, 7–20.

Tosi, H., & Tosi, D. Some correlates of role conflict and ambiguity among public school teachers. *Journal of Human Relations*, 1970, **18**, 1068–1075.

Vroom, V. *Some personality determinants of the effects of participation*. Englewood Cliffs, N.J.: Prentice-Hall, 1963.

(Received June 1, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 4, 500-503

STRESS AND PERFORMANCE DURING AND AFTER OFFICER CANDIDATE SCHOOL

J. RICHARD JENNINGS,[1] R. M. ROSE,[2] AND LEO KREUZ[3]

Walter Reed Army Institute of Research, Washington, D.C.

Military performance during and after Officer Candidate School (OCS) was related to interview and questionnaire measures of stress and personality. A company of candidates was studied early in training, just prior to graduation, and after six months on duty as officers. Psychiatric interview ratings of stress and change in coping style related to class standing in OCS but not to subsequent military performance. Personality and maturational variables were related to performance after OCS. Motivational distortion by candidates and initial impression formation by superior officers appeared to be factors in determining OCS class standing.

Officer Candidate School (OCS) attempts to place an individual under physical and psychological stress akin to that found in combat (U.S. Department of the Army, 1969). Candidates who successfully complete training under these conditions are commissioned second lieutenants.

The current study examined the candidate's reactions to the stress of OCS and his personal coping resources. Three psychological measurement approaches were used: (*a*) a psychiatric interview assessing "here and now" feelings and functional impairments during OCS; (*b*) a semi-projective test designed to measure reactions to the specific stressors of OCS and also to measure the individual's coping repertory vis-à-vis these stressors; (*c*) a set of standard paper-and-pencil tests of personality traits.

The psychological measures were related first to leadership rank (i.e., class standing) during an early, highly stressful period; second, to final OCS class standing; and third, to an assessment of the candidate's performance after approximately six months on his first permanent duty assignment.

METHOD

Subjects and Design

Subjects consisted of an entire OCS company of 69 men at Fort Benning, Georgia, 12 of whom did not complete the course and who were dropped from the sample.

The first assessment was performed during a high-stress period (Week 3 within the 23-week program). Collection of test scores, interview data, and biological sampling (see Kreuz, Rose, & Jennings, 1972) took place over three days within this third week. An identical three-day period of data collection was conducted during a minimal stress period (Week 22). A final assessment was made approximately six months after the candidates reported to their first permanent military assignment.

Performance Measures

Leadership rank (class standing within OCS) was the basic measure of the candidates' ability to function efficiently under stress. The ranks are assigned through a combination of scores on academic military science subjects, physical training scores and, most importantly, ratings by their immediate military superiors (tactical officers) and their fellow candidates. A ranking of the candidates is performed approximately every four weeks. Ranks are assigned within a platoon until the final ranking period, when a company rank is assigned.

In addition to these ranks, the physical training scores of the men were included as a separate variable, and an index of performance requirements during Week 3 of OCS was created. Termed "chain of command" points, this score reflected the student command positions which a candidate held during Week 3. Thus, for example,

---

[1] The assistance of Jim Struthers in the development and administration of the telephone follow-up interview and of Al Bardoni in the scoring of the coping scale is gratefully acknowledged. Thanks are also due the Officer Candidates and Staff of OCS at Fort Benning and Fort Belvoir for their cooperation and assistance.

Requests for reprints should be sent to J. Richard Jennings, Department of Experimental Psychophysiology, Forest Glen, Building 189, Walter Reed Army Institute of Research, Washington, D.C. 20012.

[2] Now at Department of Psychosomatic Medicine, Boston University School of Medicine.

[3] Now at the Washington Hospital Center, Washington, D.C.

a student platoon leader received 4 points and a squad leader 2 points.

Performance on active duty was assessed from a structured telephone interview with the former candidate's immediate supervisor. The supervisor assigned a score ranging from 1 to 7 for performance within each of five areas. The areas were (*a*) punctuality in completion of work, (*b*) completeness of the work, (*c*) quality of the work, (*d*) speed of learning new skills, and (*e*) budgeting and organizing time.

Psychiatric Interview

The three investigators as a group interviewed each candidate for approximately 20 minutes during both high- and low-stress measurement periods. The format of the interview was structured so that each candidate received the same questions with similar wording; however, the questions were frequently reordered, and specific additional questions added. After each interview, each investigator independently scored the individual at 1 to 10 in each of five categories: (*a*) performance and functional impairment, (*b*) anxiety, emotional response, arousal, (*c*) threat to self-esteem, (*d*) impulsivity, and (*e*) global stress. In addition, a report of sexual activity and sexual maturation was taken.

The average correlation between the global stress ratings of the three interviewers was .80 during the high-stress period. This figure dropped somewhat during the less stressful period due to the uniformly low ratings assigned to the candidates. Averaged ratings were used for the data analysis.

Stress Reaction Scale

A semiprojective test of reactions to seven specific stress situations encountered in OCS was developed according to the techniques of Goldfried and D'Zurilla (1969). A candidate was asked to put himself into a situation, numerically rate his emotional reaction, describe what he would do about the situation, and predict the outcome of the situation. The following is an item from the scale:

> After a few days of academic classes in OCS, your tactical officer heard that your class was lacking in attentiveness and military bearing. As a result, your class spent an hour low crawling and running. The next day you try especially hard to maintain a military bearing and address the instructor as you are supposed to. Many of your classmates, however, do not seem to modify their classroom manner at all. You are concerned that you will be punished as a class again.

The answers were scored by two judges for coping failure, solution inadequacy, and unfavorability of the predicted outcomes. Scores from 1 to 5 were assigned; higher scores indicated greater stress. Interjudge correlations were .77 for coping failure, .75 for solution inadequacy, .76 for unfavorability. Averaged ratings were used for the analyses.

The 7-item stress reaction scale was administered during the high-stress period at Fort Benning. A shortened 3-item form, based on the candidate's individual choice of most relevant situations, was administered during the low-stress period. The tests were administered in a group setting, and candidates wrote out their answers. In addition to the basic scores for each category averaged over items, difference scores were created by subtracting low-stress-period item scores from the relevant high-stress item scores.

Paper-and-pencil tests

During the early, stressful period, the Marlowe-Crowne Social Desirability Scale, the Buss-Dirkey Test of Aggression, and IPAT (Institute for Personality and Ability Testing) Anxiety scale were administered in a single testing session. During the less stressful period the 16 PF (Personality Factor) test, Form A, 1962 edition, was administered. After this testing session, the candidates rated each week of the OCS program for the level of stress.

RESULTS AND DISCUSSION

Difference between the High- and Low-Stress Periods

Mean differences between high- and low-stress periods for variables spanning this interval showed consistently the stressfulness of Week 3 of OCS as compared to Week 22.[4] The stress reaction scale results indicate a decrease in the failure to cope score ($t = 4.69$, $p < .001$), but no change in emotional reactions or assessments of the outcome of situations.

The overall group results for the 16 PF scales may be relevant as a description of the men at the end of OCS. In comparison to the general population norms, the men scored significantly higher than the norms on five scales ($ts > 2.37$, $df = 1183$, $p < .05$). The men's scores were substantially higher than were the norms for Intelligence (Scale B), Dominance (Scale E), and Surgency (Scale F). (Surgency is defined as enthusiastic, heedless, happy-go-lucky.) The men scored slightly higher than the norms for Ego

---

[4] For more complete tables, readers may write the first author.

TABLE 1

CORRELATIONS BETWEEN OCS (OFFICER CANDIDATE SCHOOL) PERFORMANCE AND SELECTED VARIABLES

Variable	Performance scores			
	Platoon rank: Week 4	Final company rank: Week 23	Physical training	Follow-up performance
Psychiatric interview[a]				
Functional impairment	−.46**	−.33*	−.26*	−.11
Anxiety	−.28*	−.16	−.15	−.10
Threat	−.35**	−.28*	−.14	−.10
Global stress	−.43**	−.37**	−.23	−.16
Functional impairment: Week 3–Week 22	−.33*	−.17	−.17	.01
Stress Reaction Scale[a]				
Emotion	.02	.08	−.24	.15
Coping Failure	−.12	.05	−.09	.07
Solution Inadequacy	−.03	.05	−.12	.06
Unfavorability	−.07	.04	−.02	.03
Coping: Week 3–Week 22	.25	.29*	.22	.19
16 PF scales[b]				
Shy vs. Socially Bold (Scale H)	.23	.27*	.15	−.06
Practical vs. Imaginative (Scale M)	.21	−.09	.02	.31*
Forthright vs. Shrewd (Scale N)	−.25	−.26*	−.14	−.33*
Motivation Distortion: low vs. high	.22	.32*	.13	−.01
Puberty and sex related scores				
Onset of puberty	.13	.16	−.13	.34**
Onset of masturbation	−.05	−.02	−.12	.28*
Performance				
Platoon rank	—	.70**	.50**	.09
Company rank	—	—	.46**	.09
Follow-up performance	—	—	.19	—
Chain of command: Week 3	−.29*	−.22	−.49**	.04

Note. Parallel significant correlations with performance variables were found for all the psychiatric interview change scores. Correlations with leadership ranks did not change when the physical training score was partialed out. $N = 58$; $N = 57$ for follow-up score. Of the 196 correlations from the total 4 × 49 matrix, 15% of independently assessed variables were significant.

* $p < .05$ ($p = .05$ at $r = .26$).

** $p < .01$ ($p = .01$ at $r = .34$).

[a] High scores indicate stress.

[b] High score indicates possession of second trait.

Strength (Scale C) and Self-sufficiency or Non-group-dependent (Scale Q_2).

Variables Relating to Performance during and after OCS

Table 1 presents correlations of variables that related significantly to at least one of the performance measures. A number of other variables are added to give a balanced picture of the results.[5]

The best predictors of rank during OCS were the variables from the psychiatric interview. Performance after OCS, however, was not related to these variables. All interview data showed similar correlations and probably reflected a general assessment of amount of disturbance the candidate expressed. As expected, during OCS greater expressed stress was associated with poor class standing. The psychiatric interview did not pinpoint particularly debilitating forms of psychological stress. Holmen, Katter, Jones, and Richardson (1956) also reported general ratings which predicted OCS performance but were unable to identify specific predictors.

---

[5] A correlation matrix containing all of the variables may be obtained by writing the first author.

The coping failure score of the stress reaction scale showed a negative relation to OCS rank, but only after the men had an opportunity to learn how to cope within the system. That is, the coping failure score from Week 3 did not relate to performance, but the difference score did. Candidates who shifted their answers toward an active, aggressive coping style attained better class standings than those who did not shift. Using the Job Analysis and Interest Measure (JAIM), Petersen (1971) and Petersen and Lippitt (1968) found similar shifts of behavioral style within OCS. Petersen also found that behavioral style was related to success in OCS.

Four scales from the 16 PF test related to performance. In terms of the scale descriptions of the 16 PF test, the venturesome, socially bold candidate ranked better at Week 23 than did the shy, reserved candidate; and the forthright, natural candidate did better than the shrewd, calculating candidate both at Week 23 and after 6 months on duty. Imaginative and inner-directed candidates were rated high by supervisors after 6 months on duty, although this scale was not related to OCS class standing. The fourth relationship involves motivational distortion, a lie scale. The scale is keyed on shifts in response patterns when the test is given in a job application setting versus a neutral setting. The candidates showing motivation distortion tended to have better ranks than the other candidates; that is, candidates who were concerned with "looking good" were rated as doing better.

The fairly marked relationships between later onset of puberty and masturbation and higher performance ratings after OCS were unanticipated. The result may indicate the value placed on his officer status by the late-maturing individual (cf. Jones, 1957).

Within OCS, rank at Week 4 was highly correlated to company rank (Week 23), and both showed a substantial relation to the physical training score. The correlation between rank at Week 4 and final rank reflect the relative fixing of the top and bottom third of the ranks by Week 4. The correlations of the chain of command points indicate a selection factor in the placement of men on the chain of command. Men doing relatively poorly in rank and physical training performance were more likely than others to be placed in stressful student leadership roles during Week 3. In short, the performance correlations within OCS suggest to some extent that men are labeled early in OCS as "leaders" or "duds" and then treated accordingly.

The OCS performance measures did not correlate with the rating of subsequent military performance. Given that the goal of OCS is to produce combat leaders, two facts are relevant to the absence of correlation between OCS and later performance. First, the ratings of permanent duty performance emphasized efficiency on the job rather than more abstract leadership qualities. Second, 94% of the men were serving in technical and administrative leadership positions rather than field or combat positions. A study performed during the Korean conflict (Katter & Holmen, 1954), however, found correlations (.15 to .20) similar to the current results between OCS ranks and combat performance ratings.

In summary, candidates who succeeded at OCS tended to be forthright, socially bold individuals who adopted active coping styles. They expressed relatively few problems due to stress, but may have adopted some alien attitudes in order to succeed. Their success may in part be due to making a good impression on officers during the first few days of OCS.

REFERENCES

Goldfried, M. R., & D'Zurilla, T. J. A behavior-analytic model for assessing competence. In C. D. Spielberger (Ed.), *Current topics in clinical and community psychology*. New York: Academic Press, 1969.

Holmen, M. G., Katter, R. V., Jones, A. M., & Richardson, I. F. *An assessment program for OCS applicants*. (Tech. Rep. 26) Alexandria, Va.: Human Resources Research Organization, 1956. (NTIS No. AD-91 213).

Jones, M. C. The later career of boys who were early- or late-maturing. *Child Development*, 1957, **28**, 113–128.

Katter, R. V., & Holmen, M. G. *Infantry OCS evaluations and combat performance*. (Tech. Rep. 8) Alexandria, Va.: Human Resources Research Organization, 1954. (NTIS No. AD-39 552).

Kreuz, L., Rose, R. M., & Jennings, J. R. Suppression of plasma testosterone levels and psychological stress. *Archives of General Psychiatry*, 1972, **26**, 479–482.

Petersen, P. B. An investigation of the effect of training. Unpublished doctoral dissertation, George Washington University, 1971.

Petersen, P. B., & Lippitt, G. L. Comparison of behavior styles between entering and graduating students in Officer Candidate School. *Journal of Applied Psychology*, 1968, **52**, 66–70.

U.S. Department of the Army. *The OCS story*. Washington, D.C.: Department of the Army, Headquarters, October 1969. (Pamphlet)

(Received June 25, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 4, 504-506

A PERFORMANCE CORRELATE OF PERSONAL CONTROL BELIEFS IN AN ORGANIZATIONAL CONTEXT

W. J. HEISLER[1]

University of Notre Dame

The internal-external control beliefs of 175 employees of a government agency were compared to an index of personal effectiveness composed of five factors indicative of success within a governmental occupational context: number of promotions, salary increases, and awards received; current salary; and grade differential. Employees with a greater belief in external control (Rotter scale) reported significantly lower personal effectiveness than those with a lesser belief in external control. When organizational rewards were perceived by agency members as skill-determined, the relationship between environmental control beliefs and personal effectiveness increased, while the relationship was nearly zero when agency rewards were seen as chance-related.

Over the past decade, a significant amount of research has examined what several investigators consider a major variable in behavior theory: the generalized beliefs that individuals have about their ability to influence or control their environment.[2] Although a majority of the studies in this area have focused upon the effects of such beliefs upon individuals' efforts to acquire information and engage in social influence activities (Davis & Phares, 1967; Gore & Rotter, 1963; Seeman & Evans, 1962), more recently a number of investigators have been concerned with the relationship between internal–external control beliefs and what might be termed "personal effectiveness" (Lao, 1970; Phares, 1965).

While these latter studies have tended to suggest a modest, but significant, relationship between personal effectiveness and environmental control beliefs, much of the evidence has been obtained using samples of college students in either academic or relatively simple, short-term social influence situations. It was the purpose of this study to ascertain whether a relationship between environmental control beliefs and personal effectiveness could be substantiated within a more complex organizational context. Were such associative properties to exist, it was hypothesized that organizational members with greater beliefs in their ability to control their environment would exhibit significantly greater personal effectiveness or success (as indicated by the receipt of greater organizational rewards) than would organizational members with lesser beliefs in their environmental control abilities.

METHOD

The study was carried out in a state office of a United States governmental agency which provides technical assistance to various groups and individuals in land and water resource management. Since agency personnel were geographically dispersed throughout the state, questionnaires with measures of the various variables in this study were mailed to all employees at their job address with a provision for direct mail return to the researcher. Returns were received from 196 personnel, 75% of those surveyed, although only 175 were substantively complete enough for data analysis. Because of the nature of the occupational field, nearly all respondents were male with an age range of 20 to 60, a mean age of 41.2, and a standard deviation of 11.5. Approximately 30% of the respondents held managerial positions; the remaining 70% were technical specialists and staff personnel. Of the respondents, 10% reported having attended graduate courses; 60% reported holding a four-year college degree; the remainder had at least a high school education.

The Internal–External (I–E) scale developed by Rotter (1966) was used as a measure of each individual's generalized expectancies or beliefs as to whether events are contingent upon his own behavior. For the respondents in the study, scores ranged from 0 (the consistent belief that individuals can influence their environment) to 20 (the consistent belief that the locus of reward is external to the individual, i.e., determined by luck, chance, powerful others, etc.) with a mean of 7.6 and a standard deviation of 4.3.[3]

---

[1] Requests for reprints should be sent to William J. Heisler, Department of Management, College of Business Administration, University of Notre Dame, Notre Dame, Indiana 46556.

[2] An excellent summary of this research can be found in Rotter, Seeman, and Liverant (1962) and Rotter (1966).

[3] While this mean is somewhat more internal than a number of the general norms reported by Rotter (1966), it is consistent with Gemmill and Heisler's (1972) previous results for managerial personnel.

The measure of personal effectiveness was a locally constructed index incorporating a number of items generally considered to be global indicators of performance effectiveness or success within a governmental occupational context: number of promotions, within-grade salary increases, and awards received; current salary; and grade differential (the difference between the employee's current general schedule (GS) grade and level and his starting GS grade and level with the agency). Responses for each item were recoded as nearly as possible into five equal categories and assigned weights of one to five. A total score for each respondent was then obtained by summing across all items in the index. Scores on the measure ranged from 5 (low personal effectiveness) to 22 (high personal effectiveness) with a mean of 13.3 and a standard deviation of 4.5. The internal consistency of the measure was .69.

RESULTS

The personal effectiveness index was found to be significantly correlated with the personal control beliefs of agency members ($r = -.25$, $p \leq .01$), indicating that those members who expressed a greater belief in their ability to exert influence over their environment reported experiencing greater personal effectiveness. The personal effectiveness index, which is cumulative in nature, was also found to be a function of the respondent's age ($r = .41$, $p \leq .001$), education ($r = .20$, $p \leq .05$), and length of service with the agency ($r = .58$, $p \leq .001$). Therefore, third-order partial correlation coefficients were calculated to eliminate statistically the effects of these factors. The resulting degree of association between agency members' environmental control beliefs and personal effectiveness decreased somewhat ($r = -.22$), but remained significant at the .01 level.

Since prior research also indicated that a person's knowledge of the chance or skill nature of the specific task situation which he faces appears to moderate the effects of his more general personal control beliefs (Davis & Phares, 1967; Rotter & Mulry, 1965), a measure of employee perceptions of the chance or skill nature of the agency reward structure was also incorporated in the questionnaire. More specifically, members were asked to indicate which of several factors (e.g., ability and effort, knowing the right people, being in the right place at the right time, etc.) they saw as being most important in influencing promotion and salary decisions within the agency. Responses were classified on a priori grounds into the dichotomous categories of skill and chance.

As hypothesized, the difference in personal effectiveness between internals and externals was greater under skill than under chance perceptions, with internals experiencing greater personal effectiveness. For those agency members who saw organizational rewards as skill-determined ($n = 93$), the association between environmental control beliefs and ratings of personal effectiveness was significant at the .001 level ($r = -.34$). The association decreased to $r = -.27$ when the effects of age, education, and time with the agency were controlled for, but remained significant at the .01 level. However, for those employees who saw agency rewards as chance-related ($n = 80$), the relationship between environmental control beliefs and personal effectiveness was nearly zero ($r = .01$), with a third-order partial $r = -.08$.

SUMMARY

The results of this study support the hypothesis that organization members with a greater external control orientation experience significantly less personal effectiveness (as indicated by the number of promotions, salary increases, awards, etc., that they receive) than do members with a lesser external orientation. This finding appears consistent with the results of a study of Gemmill and Heisler (1972) in which managers with more external control beliefs reported experiencing significantly greater job strain and less job satisfaction than did managers with a lesser external orientation. To the extent that externality is associated with lower levels of organizational reward, one might expect the holder of such beliefs concomitantly to feel less satisfied with his job (particularly with his job progress) and more bothered by various elements in the job situation (e.g., feeling unable to influence superiors' actions and decisions, not knowing how a superior evaluates his performance, not knowing what opportunities exist for promotion, etc.).

The results further support prior conclusions concerning situational saliency in the effects of personal control beliefs. When promotions, salary increases, and awards were perceived by agency members as skill-determined, the differences in personal effectiveness between internals and externals increased in significance. However, when organizational rewards were perceived as chance-related, the internal–external control dimension lost its salience as a predictor of personal effectiveness.

Although the findings reported in this study are statistically significant at or below the .01 level and remain so even when subjected to partial correlational analysis, the strength of the associations is somewhat low by a number of contemporary standards. Nevertheless, when viewed as resultants of a field study within a complex organizational context, their associative properties remain of substantial interest and import.

REFERENCES

Davis, W. L., & Phares, E. J. Internal–External control as a determinant of information-seeking in a social influence situation. *Journal of Personality,* 1967, **35**, 547–561.

Gemmill, G. R., & Heisler, W. J. Fatalism as a factor in managerial job satisfaction, job strain, and mobility. *Personnel Psychology,* 1972, **25**, 241–250.

Gore, M., & Rotter, J. B. A personality correlate of social action. *Journal of Personality,* 1963, **31**, 58–64.

Lao, R. C. Internal–External control and competent and innovative behavior among Negro college students. *Journal of Personality and Social Psychology,* 1970, **14**, 263–270.

Phares, E. J. Internal–External control as a determinant of the amount of social influence exerted. *Journal of Personality and Social Psychology,* 1965, **2**, 642–647.

Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs,* 1966, 80(1, Whole No. 609).

Rotter, J. B., & Mulry, R. C. Internal versus external control of reinforcement and decision time. *Journal of Personality and Social Psychology,* 1965, **2**, 598–604.

Rotter, J. B., Seeman, M., & Liverant, S. Internal versus external control of reinforcements: A major variable in behavior theory. In N. F. Washburne (Ed.), *Decisions, values, and groups.* London: Pergamon, 1962.

Seeman, M., & Evans, J. W. Alienation and learning in a hospital setting. *American Sociological Review,* 1962, **27**, 772–781.

(Received July 19, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 4, 507–508

COMMENT ON LOCKE'S SATISFIERS AND DISSATISFIERS AMONG WHITE-COLLAR AND BLUE-COLLAR EMPLOYEES

ROBERT G. HADLEY[1]

California State University, Los Angeles

An article published by Locke in the August 1973 issue of the *Journal of Applied Psychology* offered an improved classification system for occupational satisfiers and dissatisfiers, distinguishing between agent and event categories. However, some methodological problems appear to limit the conclusions that can be drawn from Locke's study. An alternative classification system, taking into account the multiple determination of some events, is recommended.

Locke (1973) is to be commended for offering a long-needed improvement over previous systems of classifying occupational satisfiers and dissatisfiers. Of particular importance is the distinction between *agent* categories and *event* categories. He is also to be commended for offering his study for publication despite acknowledged weaknesses in his sampling procedures. The reader may thus come to his own judgment regarding the confidence he wishes to place in the conclusion, given these sampling problems.

Similarly accessible to reader judgment are five of the eight statistically nonsignificant comparisons Locke offers in his Tables 2 and 3. The percentages observed for "good" and "bad" events are so similar as to require no statistical test. However, methodological problems of analysis cast doubt on the remaining three comparisons in these tables and all those comparisons in Tables 4 and 5.

A common misuse of chi-square cited by Lewis and Burke (1949) is violation of the requirement that each tally in a chi-square table be independent of the others. In Locke's study, violation of this independence requirement occurred in two ways. In Study 1, Locke elicited descriptions of four events, two "good" and two "bad" from each of his subjects. Locke then applied the chi-square statistic to fourfold contingency tables in which events were counted rather than persons. Independence, in the context of this analysis, requires that each description of an event be elicited from a different subject so that there is no basis for correlations among the incidents. In Study 1, collecting two sets of data from each subject (Table 2) and treating them as if they represented different subjects led to the type of nonindependence of observations that inflated the *N*s, and probably biased the chi-squares upward. Additionally, in the analyses of Tables 4 and 5, inflation of *N*s occurs in both studies.

The second violation of the requirement of independence occurred in the analysis of both Study 1 and Study 2 (in which each subject described one good and one bad event) involving the comparison of good and bad events (Tables 2 and 3). The analysis failed to consider possible correlation between these two event categories; if such correlation is positive, the values of chi-square would be biased downward. However, examination of the percentages in Table 3 reveals them to be so similar that the conclusions were probably not affected by any downward bias. While the inflated *N*s in Table 2 biased the chi-squares upward, and the possible correlation between the two event categories may have biased the chi-squares downward, the magnitude of these possible biases is unknown, and consequently so is the extent to which they may compensate each other.

McNemar's (1969, p. 261) test of correlated proportions could appropriately be applied to compare the *first* good and bad incidents elicited from each subject. This procedure would require discarding some of the data of Study 1, but would use all the data of Study 2. An advantage of the McNemar test is that it capitalizes on possible correlation between good and bad events, a valuable feature of the repeated-measures design. An alternative for Study 1, which would use all the data, would be to classify each subject according to whether the number of motivator-determined good events he described (2, 1, or 0) was greater than, equal to, or less than his number of motivator-determined bad events. The sign test (Siegel, 1956) could then be applied. This same procedure could again be followed for the task–nontask analysis.

In addition to the problem of inflated *N*s, Table 4 reveals a second item on Lewis and Burke's

---

[1] Requests for reprints should be sent to Robert G. Hadley, Coordinator, Rehabilitation Counseling Program, California State University, 5151 State University Drive, Los Angeles, California 90032.

(1949) list of common infractions, that is, failure to consider frequencies of nonoccurrence. It appears that the sums of the white-collar and blue-collar columns furnished the basis for the expected frequencies against which the observed frequencies were compared. The chi-square of 18.91, for example, appears to be based on comparing the observed frequencies of 90 and 32, respectively, to expected frequencies of 66 and 56, without considering nontask frequencies. Because of the problem of inflated *N*s, this reviewer did not undertake the relatively straightforward task of recalculating the chi-squares of Table 4 with frequencies of nonoccurrence included.

My own preference would be for a system of analysis which takes account of the *multiple* determination of some events. For example, if a worker reports being given more work and additional responsibility without the promotion which normally goes with the responsibility, or without the pay raise which normally goes with the increased work, four of Locke's categories are involved (amount of work, promotion, responsibility, and money). To classify the event into *one* of these categories is to throw away some possibly valuable data. I propose assigning a constant number of points (e.g., 5 or 10) to each event and instructing the judge to allocate these points to the various categories according to the importance of each category in the event. Correlation statistics could then be used to evaluate interjudge reliability of each category, and mixed-design analysis of variance methods (Lindquist, 1953) applied to comparisons of various event categories (e.g., good/bad, task/nontask) or different groups of subjects, with tests of interaction effects possible.

Pending reanalysis of the data, it is recommended that the reader hold in abeyance any conclusions based on Locke's chi-square values. Tentative inferences based on inspection of the obtained frequencies, remembering their possible intercorrelations, are not compromised by the methodological problems cited previously.

REFERENCES

Lewis, D., & Burke, C. J. The use and misuse of the chi-square test. *Psychological Bulletin*, 1949, 46, 433–489.

Lindquist, E. F. *Design and analysis of experiments in psychology and education*. New York: Houghton Mifflin, 1953.

Locke, E. A. Satisfiers and dissatisfiers among white-collar and blue-collar employees. *Journal of Applied Psychology*, 1973, 58, 67–76.

McNemar, Q. *Psychological statistics*. (4th ed.) New York: Wiley, 1969.

Siegel, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.

(Received October 23, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 4, 509–510

ACHIEVEMENT MOTIVATION AMONG MANAGERS OF SMALL BUSINESSES, CORPORATION MANAGERS, AND BUSINESS STUDENTS

JOHN C. SOUTH[1]

School of Business and Administration, Duquesne University

A questionnaire measure of achievement motivation tested on college students in the Netherlands was administered to business college students, small business managers, and corporation managers. In addition, category scores were developed that reflect the more specific components of the n Ach (need for achievement) construct. Item analyses and Kuder-Richardson 20 estimates were adequate. Intercorrelations among the category scores were significantly lower for the student sample, which suggests the n Ach construct may be less differentiated among older managers. Managers differed significantly from business students on several components of n Ach.

The measurement of achievement motivation (n Ach—need for achievement) has both theoretical and practical importance. In a recent article, Hermans (1970) reviewed efforts to measure achievement motivation and presented the results (and items) of an empirically developed questionnaire designed to measure n Ach of college students in the Netherlands.

Hermans also presented a taxonomy of achievement motivation concepts. Ten categories of n Ach, all supported by prior research, served as a framework for developing the questionnaire. The categories are as follows: aspiration level, risk-taking behavior, upward mobility, persistence, task tension, time perception, time perspective, partner choice, recognition behavior, and achievement behavior.

For further empirical study of the questionnaire, nine categories were examined (risk-taking behavior was excluded), and the concept of a n Ach questionnaire measure was extended to the realm of business management. Specifically, the purposes of this study were (*a*) to examine the achievement motivation questionnaire using subjects employed in the real world of business; (*b*) to examine the taxonomy of the achievement motivation construct used by Hermans by intercorrelating the category scores; and (*c*) to compare the n Ach motivation levels of managers working in large corporations with those of manager–owners of small enterprises and undergraduate students majoring in business administration.

METHOD

Questionnaire and scoring procedures. The n Ach questionnaire consists of 29 items with four to six alternatives. Nine category scores were obtained by summing over those items comprising each category according to Hermans' item-taxonomy designations. The total score consisted of all 29 items. Two demographic items measured age and years of employment with present organization.

Samples and sampling procedure. A questionnaire and cover letter were mailed to 120 businessmen selected from the local Small Businessmen Association files with the only restriction that the organization employ 50 or fewer people. Forty-one usable (34%) questionnaires were returned. Personal visits to the central personnel offices of large nationally known corporations were made to collect the corporation manager data. Five major corporations participated, and 80 questionnaires were distributed among lower-middle managers with 50 or fewer subordinates. Thirty-six usable forms (45%) were obtained. The undergraduate sample consisted of business students in an upper-level, required management game course. The entire class (45 males) completed usable questionnaires. All questionnaires were completed anonymously.

RESULTS

Item analyses. The item-total score correlations (Pearson) were computed for each sample.[2]

---

[1] The author wishes to thank J. Thomas Phlegar of Humble Oil and Refining Company for his efforts in collecting the data.

Inquiries concerning this article should be sent to John C. South, School of Business and Administration, Duquesne University, Pittsburgh, Pennsylvania 15219.

[2] Detailed tables of item-total score correlations by samples and comparisons of these correlatoins with the item-cluster correlations reported by Hermans can be obtained from the author.

Only one item-total score correlation was negative, item 20 in the student sample analysis ($r = -.24$), and no category-related pattern of nonsignificant correlations emerged. The item correlations with either work experience or age were within chance expectations; none correlated consistently across samples.

Category intercorrelations. A cell by cell comparison using the r to z transformation revealed that the intercorrelation matrices for the small business manager and corporation manager samples did not differ significantly at the .05 level or beyond; both manager matrices did differ significantly from the business student matrix. Based on these results (and the t-test results), the small business manager and corporation manager samples were combined.

The intercorrelations were distinctly higher ($\bar{r} = .32$, signs disregarded) for the combined manager sample than for the business student sample ($\bar{r} = .17$, signs disregarded). None of the categories nor total score correlated significantly with years of work experience or age.

Reliabilities. The total score Kuder-Richardson 20 reliability of .67 for the business student sample is considerably below the Kuder-Richardson 20 reliabilities of .85 and .88 reported by Hermans. However, the Kuder-Richardson 20 reliabilities of .79 and .82 for the small business managers and corporation managers, respectively, are comparable.

The Kuder-Richardson 20 reliabilities of the category scores ranged from .01 to .69. In general, the reliabilities were highest for the corporation manager sample.

Comparison of means by sample. The business student sample had the lowest mean scores among the samples which suggest less persistence (student versus small business manager $t = 2.43$, $p < .05$; student versus corporation manager $t = 3.03$, $p < .01$; two-tailed tests); different partner-choice tendencies (student versus small business manager $t = 1.91$, $p < .07$; student versus corporation manager $t = 2.53$, $p < .05$; two-tailed tests); lower-task tension or less tendency to resume a task when interrupted (student versus corporation manager $t = 2.53$, $p < .05$; two-tailed test) and less seeking of recognition by performing well than do managers in corporations (student versus corporation manager $t = 2.21$, $p < .05$; two-tailed test). Similarly, the total score n Ach mean was lowest for the student sample (student versus corporation manager $t = 2.64$, $p < .01$; two-tailed test). None of the sample means differed significantly between the small business managers and corporation managers, although the corporation managers had higher mean scores in seven of the nine categories compared.

DISCUSSION

Although Hermans' questionnaire was empirically developed using college students in the Netherlands, it appears to function somewhat more adequately with older, experienced managers than with business college students in the United States. Overall, however, these results suggest further use of Hermans' instrument.

Although studies by Mitchell (1961) and Lynn (1969) yielded conflicting results, the development and use of more specific category scores are consistent with the view that the achievement motivation construct is a composite of concepts or components. Among more experienced managers and adults in the world of work, the separate aspects of achievement may have fused together as suggested by the category intercorrelation results.

The pattern of slightly higher, but not statistically significant, mean scores on most of the nine categories, specific components of achievement motivation by lower-level managers in large corporations compared to manager-owners of small businesses conforms with prior research using a global n Ach score. McClelland (1961), in a study of 26 small company managers, found that executives in large organizations had slightly higher, but not statistically significant, levels of n Ach. He found no evidence that ownership is important to subjects with high n Achievement.

In terms of application, the use of n Ach category scores may increase the utility of the questionnaire. Such scores, if sufficiently reliable, should increase the understanding of an individual's achievement motivation as well as identify aspects of n Ach for which different groups of subjects may have similar scores and/or patterns of scores. Use of the total score would mask such similarities and differences.

REFERENCES

Hermans, H. J. M. A questionnaire measure of achievement motivation. *Journal of Applied Psychology*, 1970, **54**, 353–363.

Lynn, R. An achievement motivation questionnaire. *British Journal of Psychology*, 1969, **60**, 529–534.

McClelland, D. C. *The achieving society*. Princeton, N.J.: Van Nostrand, 1961.

McClelland, D. C. N Achievement and entrepreneurship: A longitudinal study. *Journal of Personality and Social Psychology*, 1965, **1**, 389–392.

Mitchell, J. V., Jr. An analysis of the factorial dimensions of the achievement construct. *Journal of Educational Psychology*, 1961, **52**, 179–187.

(Received June 7, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 4, 511–512

EFFECTS OF APPLICANT'S SEX AND DIFFICULTY OF JOB ON EVALUATIONS OF CANDIDATES FOR MANAGERIAL POSITIONS

BENSON ROSEN[1] AND THOMAS H. JERDEE

University of North Carolina

This experiment investigated the influence of sex role stereotypes on evaluations of candidates for managerial positions. Subjects were 235 undergraduate business students. Findings confirmed that sex role stereotypes influence evaluations of applicants and selection decisions. Male applicants were accepted more frequently than were equally qualified females, and were evaluated more favorably on general suitability, potential for long service and potential for fitting in well in the organization. Also supported was the hypothesis that job demands have differential effects on evaluations of male and female applicants. Lowest acceptance rates and poorest evaluations were given female applicants for "demanding" managerial positions.

Two previous studies by the present authors have demonstrated that sex role stereotypes influence evaluations of males and females across a broad range of organizational situations (Rosen & Jerdee, 1973, 1974). Results of the second study indicated a tendency for male administrators to discriminate against women in personnel decisions involving promotion, development, and supervision.

Discriminatory behavior against women seems based on a set of attitudes which depict women as limited in the toughness, stability, creativity, and judgment required to meet the demands of high-level managerial positions. The present study investigates the extent to which these attitudes are manifested in the evaluation of female applicants for a variety of managerial jobs. We hypothesized that female applicants for managerial jobs are evaluated as less acceptable than are male applicants with identical qualifications. We hypothesized further that discrimination against female applicants is more severe when challenging or demanding activities are required by the managerial position.

METHODS

Subjects

Subjects were 235 male undergraduate business students enrolled in an organizational behavior course at the University of North Carolina during the fall semester of 1972.

Procedure

Experimental materials were presented in the form of a personnel selection exercise. Subjects were instructed to take the role of a consultant to a clothing manufacturer who had to fill four executive positions as a result of a recent expansion. Subjects were asked to consider a set of job descriptions and to evaluate an applicant for each position.

Independent Variables

The two independent variables were sex of the applicant and job demands. A standardized application form provided subjects with information about each job applicant. The personal data consisted of name, age, place of birth, marital status, education, a description of relevant work history, and interviewer's comments. An attempt was made to construct these applicant descriptions so that the candidates would have an even chance of being accepted or rejected, since we were interested in subjects' reactions to a typical or average candidate. Two versions of each application form were prepared for each of the four job openings; only the name of the candidate was changed to manipulate the sex variable. For example, for one position some subjects considered a Mr. Phillip Lewis, while others considered a Mrs. Phyllis Lewis with identical background characteristics.

Two versions of job requirements were created for each of the four organizational positions (test lab manager, purchasing manager, personnel manager, and financial manager). One version included "demanding" job requirements that required aggressive interpersonal behavior and decisive managerial action. The other version consisted of "routine" behaviors that required clerical accuracy and dependable performance.

For each of the four jobs, each subject read only one version of the job requirements—demanding or routine—and reviewed information about only one applicant—male or female. The procedure can best be conceptualized as an incomplete $2 \times 2 \times 4$ repeated-measures design embedded in the personnel selection exercise. Each subject responded to each of the experimental conditions, one condition for each of the four jobs.

---

[1] Requests for reprints should be sent to Benson Rosen, Graduate School of Business Administration, University of North Carolina, Chapel Hill, North Carolina 27514.

Dependent Variables

Subjects indicated acceptance or rejection of each applicant and evaluated the applicant on the following scales: (*a*) overall rating, (*b*) potential for technical aspects of the job, (*c*) potential for long service to the organization, and (*d*) potential for fitting in well in the operation. Evaluations were marked on 6-point scales ranging from (1) extremely unfavorable to (6) extremely favorable.

RESULTS AND DISCUSSION

The major focus of the investigation was the influence of sex stereotypes and job demands on personnel decisions. As specific job requirements were irrelevant, we pooled the data from the four jobs.

Females were selected for managerial positions significantly less often than males (59% acceptance for females compared to 71% acceptance for males; $\chi^2 = 14.51$, $p < .01$). In addition, females were rated significantly lower than were males on all other dependent variables. Overall ratings for female applicants were lower than for males ($F = 14.02$, $p < .01$). Females were also rated lower than were males on "technical potential" ($F = 7.97$, $p < .01$), "potential for long service to the organization" ($F = 33.81$, $p < .01$), and "potential for fitting in well ($F = 17.26$, $p < .01$). These findings provide strong support for our first hypothesis that female applicants for managerial positions are evaluated as less acceptable than are male applicants with identical qualifications. Furthermore, it appears that the general tendency to discriminate against female applicants for managerial positions is based on perceptions and expectations regarding a female employee's interpersonal competence and dependability as well as her technical skill.

The second hypothesis concerned the differential influence of various job requirements on evaluations of male and female applicants for managerial positions. The relevant statistic is the Sex × Job requirement interaction.

The rate of acceptance for females was only 46% in the demanding condition versus 72% in the routine condition. The corresponding percentages for males are 65% and 76%. Thus the difference in acceptance rates between the demanding and easy job conditions is 26% for females versus only 11% for males ($Z = 2.24$, $p < .05$). There is a marked tendency to reject females for the demanding jobs.

The lowest evaluations on all of the dependent variables occur in the female applicant—"demanding" requirement condition. The Sex × Job Requirement interaction was significant for overall ratings ($F = 6.27$, $p < .05$), ratings of potential for long service ($F = 5.80$, $p < .05$), and ratings of potential for fitting in well in the organization ($F = 5.76$, $p < .05$). The findings of this study support the second hypothesis that discrimination against women applicants for managerial positions is more severe when the job requirements are demanding and challenging.

These results provide further evidence of the subtle, but pervasive influence of sex role stereotypes in situations where the issue of sex is not made salient for the decision maker. Additional research is required, however, to establish the generality of these findings beyond our sample of southern business students.

REFERENCES

Rosen, B., & Jerdee, T. H. The influence of sex-role stereotypes on the evaluation of male and female supervisory behavior. *Journal of Applied Psychology*, 1973, **57**, 44–48.

Rosen, B., & Jerdee, T. H. Influence of sex role stereotypes on personnel decisions. *Journal of Applied Psychology*, 1974, **59**, 9–14.

(Received June 14, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 4, 513-515

SPIN-OFFS FROM BEHAVIORAL EXPECTATION SCALE PROCEDURES

MILTON R. BLOOD[1]

University of California, Berkeley

Behavioral expectation scale procedures have been used to construct performance evaluations in a variety of work settings. With slight modifications the procedures can also be used to provide information to organizations for extending the domain of evaluated performance, developing training programs, and assessing agreement on and communication of organizational policy.

Considerable attention has been given in recent years to the construction of performance rating scales using the procedures developed by Smith and Kendall (1963) for preparing scales of behavioral expectations. These procedures involve three distinct steps. In the first step, behavioral episodes or critical incidents are generated (Flanagan, 1954) which illustrate job performance dimensions. In the second, or retranslation, step the behavioral incidents are reassorted into performance dimensions, and incidents are retained only if they are reliably placed in the dimension for which they were generated. The third step scales each remaining incident according to the level of performance it represents. Items receive scale values equal to their mean rating on a dimension representing the "goodness" of the behavior. Items with large variances are not retained.

Most accounts of the use of these procedures have dealt with the development of performance scales in an increasing variety of work settings (Fogli, Hulin, & Blood, 1971; Harari & Zedeck, 1973; Mass, 1965; Zedeck & Baker, 1972). Campbell, Dunnette, Arvey, and Hellervik (1973) have compared the behaviorally anchored scales with another type of rating, and Borman (1972) has explored the generalizability of such scales to raters other than those in the organizational position of persons assisting in scale construction. Behavioral expectation scaling techniques can be used to accomplish other organizational objectives in addition to the construction of performance rating instruments. For example, Landy and Guion (1970) have used the procedure to construct motivation scales.

To cite another example, the first step of the procedure has been suggested as an information gathering method to be used in the development of a job analysis (Blum & Naylor, 1968; Tiffin & McCormick, 1965). The critical incident type of information specifies concrete behaviors that must be performed by an individual on the job, whereas a traditional job analysis names the tools used, required education, and preferred worker characteristics. For job analysis purposes it is not necessary to eliminate items that could not be retranslated according to the second step of the procedure. In describing the behaviors of the job, statements need only to be descriptive of job behaviors, not necessarily classifiable into a limited set of performance dimensions.

In addition to the previously mentioned examples, this article presents four organizational benefits which can be derived from the procedures used in constructing behavioral expectation scales.

Extending the Domain of Evaluated Performance

Behavioral expectation procedures can be used to broaden the domain of performance evaluation. Borman (1972) has shown that when performance evaluation scales are developed from more than one perspective such as by organizational superiors and by peers, it is possible to get different behavioral dimensions from each perspective. His evidence indicates that not only do the separate perspectives contribute different performance dimensions to the evaluation schema, but also persons from each perspective are best able to rate performance on the scales developed from their perspectives. For example, peers may be better than superiors at developing and using a scale which measures cooperation, while superiors might be better than peers at developing and using a scale which measures quality of work. By using the scaling procedures from as many organizational perspectives as possible (e.g., superiors, peers, subordinates, clients) we may considerably broaden the domain of behavior which we can assess accurately and on which we can provide feedback to organization members. A retail clerk might conceivably be rated on different aspects of his performance by his department manager, by other clerks who work with him, and by customers he has served.

---

[1] Requests for reprints should be sent to Milton R. Blood, who is now at the College of Industrial Management, Georgia Institute of Technology, Atlanta, Georgia 30332.

Whether different organizational perspectives provide different scales may depend on the nature of the job and the visibility of the job holder to persons in the various organizational perspectives. Zedeck (1973) obtained performance dimensions that were highly similar even though they were developed at two separate levels of an organization.

Developing Training Programs

The behavioral items which are collected in a behavioral expectations scaling procedure could provide the basis for a training program. This suggestion is also made by Campbell et al. (1973). The skills to be trained are specified in terms of the actual job behaviors rather than simply the name of a skill domain. As for a job description, a training program can be developed from the entire set of behaviors collected and agreed upon, not just the behaviors listed in the final performance evaluation scales, that is, items eliminated by the second step can be used in the training development. The behavioral expectation items also allow trainers to specify unwanted behaviors that trainees should learn to avoid. A great deal of actual job performance description is contained in the behavioral examples which the procedures provide. This can be used to make training directly relevant to the job performance expected of the trainee. Trainees in a program based on these materials would learn expected behaviors, and they would learn how their performance is to be evaluated.

For example, Fogli et al. (1971) found an item for grocery checker performance which read "You can expect this checker to know the various sizes of cans—No. 303, No. 2, No. 2½." If this knowledge is expected of checkers, they should be trained to make the necessary discriminations. Also, the evaluation for this part of a grocery checker's training program is obvious from the behavioral item.

Assessing Agreement on Organizational Policy

The behavioral expectation scaling procedures provide an analysis of the clarity of organizational policy within the level on which the scales are developed. If materials for the scales come from managers, the item variances used in step three of scale construction indicate the agreement among managers on the appropriateness of certain behaviors. For scale development, items with large variances are discarded, but those may be the most important items for organizational development. Items with large variances pinpoint areas where organizational policy is unclear or in conflict. Some of the respondents consider the behavior to be good performance and others consider it to be poor performance. If it is an important behavior, some organizational decision should be made to determine whether it should be encouraged or discouraged. Then, the organization's policy regarding that behavior can be transmitted to the personnel who perform or evaluate the behavior. Large variances in Step 3 cue the organization to the potential problems of disagreement.

Assessing the Accuracy of Communication of Organizational Policy

In the behavioral expectation scaling procedures members of the organization rate the job behavior items according to whether the behavior represents effective or ineffective job performance. The scale value for an item is then established as the mean rating. Item means can be collected for the same items from members at two different organization levels to indicate how well policy has been communicated between levels. For instance, if scales are developed by managers for use on production workers, then the behavioral incidents could be rated for effectiveness (as in Step 3) by both managers and production workers. If managers agree that a particular behavior is highly desirable and production workers either don't agree or they agree that it is undesirable, then there is an obvious need to communicate to the production workers how this behavior (on which they will be evaluated) is viewed by managers; or, there is a need to enlighten the managers as to why this behavior should not be considered positively. This type of difference has been shown to exist in organizations by Zedeck (1973).

CONCLUSION

Perhaps because of its promise as a performance scaling technique, several potentials of the behavioral expectations procedures have been overlooked. They can provide us with helpful information in other areas of organizational operation. Applied research should be carried out to test these suggested additional uses.

REFERENCES

Blum, M. L., & Naylor, J. C. *Industrial psychology: Its theoretical and social foundations.* New York: Harper & Row, 1968.

Borman, W. C. The rating of individuals in organizations: An alternative approach. Unpublished doctoral dissertation, University of California, Berkeley, 197[illegible].

Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L. V. The development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology*, 1973, **57**, 15–22.

Flanagan, J. C. The critical incident technique. *Psychological Bulletin*, 1954, **51**, 327–358.

Fogli, L., Hulin, C. L., & Blood, M. R. Development of first-level, behavioral job criteria. *Journal of Applied Psychology*, 1971, **55**, 3–8.

Harari, O., & Zedeck, S. Development of behaviorally anchored scales for the evaluation of faculty teaching. *Journal of Applied Psychology*, 1973, **58**, 261–[illegible]5.

Landy, F. J., & Guion, R. M. Development of scales for the measurement of work motivation. *Organizational Behavior and Human Performance*, 1970, **5**, 93–[illegible].

Maas, J. B. Patterned scaled expectation interview: Reliability studies on a new technique. *Journal of Applied Psychology*, 1965, **49**, 431–433.

Smith, P. C., & Kendall, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, 1963, **47**, 149–155.

Tiffin, J., & McCormick, E. J. *Industrial psychology*. (5th ed.) Englewood Cliffs, N.J.: Prentice-Hall, 1965.

Zedeck, S. Behaviorally anchored rating scales: View from two organizational levels. Paper presented at the meeting of the American Psychological Association, Montreal, August 1973.

Zedeck, S., & Baker, H. T. Nursing performance as measured by behavioral expectation scales: A multitrait–multirater analysis. *Organizational Behavior and Human Performance*, 1972, **7**, 457–466.

(Received August 2, 1973)

ANNOUNCEMENT

As of June 1, 1974 the editorial office of the *Journal of Applied Psychology* moved to a new address. Please send all manuscripts and correspondence regarding manuscripts to Dr. Edwin A. Fleishman, Editor, American Institutes for Research, 3301 New Mexico Avenue, N.W., Washington, D.C. 20016.

Journal of Applied Psychology
1974, Vol. 59, No. 4, 516–518

A UTILITY ANALYSIS OF A WEIGHTED APPLICATION BLANK DESIGNED TO PREDICT TURNOVER FOR CLERICAL EMPLOYEES

RAYMOND LEE[1] AND JEROME M. BOOTH

Hennepin County Personnel, Minneapolis, Minnesota

The validity of a weighted application blank was demonstrated in predicting turnover for clerical employees. Cross-validated results showed relatively low shrinkage of the validity coefficient. The magnitude of correlation ($r_{pb} = .56$, $O_v = 49\%$) indicated that weighted application blank scores could be used as a basis of selection. A utility analysis indicated that the maximum potential savings could amount to about $250,000 during a 25-month period.

During the past several decades, a considerable amount of literature has reported on weighted application blank studies (Dornon, 1970; Dunnette, Kirchner, Erickson, & Banas, 1960; England, 1971; Kavanagh & York, 1972). This article reports the results of a replication study of such an application blank designed to reduce the turnover rate of clerical employees. Additionally, the study demonstrates how utility analysis can be used to provide indices of cost savings from the introduction of the application blank developed.

CROSS-VALIDITY STUDY

Two samples of intermediate level clerical employees who were hired on a permanent full-time basis were obtained. The sample on which item weights were developed consisted of 75 long-term (16 months and over, mean = 21.22) and 75 short-term (terminated within 6 months, mean = 3.64) employees hired during August 1969, to February 1972. The cross-validation sample consisted of 40 long-term (16 months and over, mean = 20.73), 20 intermediate-term (terminated within 7 to 15 months, mean = 10.05), and 34 short-term (terminated within 6 months, mean = 2.68) employees hired during March 1971, to March 1973.

This study was conducted following the procedure described by England (1971). Eleven scorable items were found to significantly differentiate between long- and short-term employees. After consulting with our Affirmative Action Department, two items (age and height) were dropped from the scoring system in order to reduce the possibility of adverse impact on minorities or protected groups. Using the remaining nine derived weights, all subjects in the two samples were scored, and a point-biserial correlation coefficient[2] (r_{pb}) between the weighted application blank scores and the tenure criterion, and Tilton's (1937) overlap coefficient[3] (O_v) were calculated for each sample. The obtained results were $r_{pb} = .59$ ($n = 150$, $p < .001$) with an $O_v = 46\%$ for the weighting sample, and $r_{pb} = .56$ ($n = 74$, $p < .001$) with an $O_v = 49\%$ for the cross-validation sample.

Obviously, the magnitude of the r_{pb} validity coefficients held up very well when the scoring system developed for the weighting sample was applied to the cross-validation sample. When the intermediate-term group was included in the cross-validation sample, a highly significant product-moment correlation coefficient was obtained, $r = .56$ ($n = 94$, $p < .001$). Therefore, it was evident that the weighted application blank scores predicted the criterion of tenure and could be used in making selection decisions. Although the tenure criterion is specifically mentioned in the Equal Employment Opportunity Commission Guidelines (1970) as an acceptable measure of job performance, the utility in dollars or business necessity of the weighted application blank had to be estimated before the blank's use was justified.

---

[1] Requests for reprints should be sent to Raymond Lee, Psychologist, Personnel, Hennepin County, 605 Fourth Avenue South, Minneapolis, Minnesota 55415.

[2] Some researchers have criticized a high point-biserial correlation coefficient in weighted application blank studies as being an overestimate of the relationship because of the extreme dichotomy of the tenure criterion. The inclusion of an intermediate term group in the cross-validation sample revealed a highly significant product-moment correlation coefficient between the tenure criterion and the weighted application blank scores ($r = .56$, $n = 94$, $p < .001$).

[3] The overlap coefficients demonstrate good predictive efficiency of the weighted application blank. Dunnette (1966) states, 'Values between 50% and 75% denote fairly good separation between the two groups and usually denote a practically useful relationship.'

UTILITY ANALYSIS

Berkson (1947) defined "utility" as the proportion of unsuccessful applicants eliminated, and "cost" as the proportion of successful applicants rejected. While some researchers, notably Cronbach and Gleser (1965), state that the rejection of applicants of good quality does not decrease the output of the institution, Dunnette (1966) argues that there are many instances in industry where the rejection of successful executives will be very costly. For a relatively low-level job, however, most psychologists agree that it is not meaningful to consider the cost of rejecting long-term clerical employees. Therefore, the purpose of this section was to evaluate the utility of adding a weighted application blank to the existing selection procedure considering the cost of eliminating unsuccessful applicants without considering the cost of rejecting a potential long-term employee.

The cost of terminating is considered to be the total cost of hiring, training, and orienting a new clerical employee to the job. A questionnaire distributed to clerical supervisors provided information necessary to this analysis of termination costs. The response to a question relating to the average length of time required for a new employee to become fully functioning at an average level of proficiency or to a level requiring only normal supervision was 5 months. The response to another question relating to average percentage of time spent in supervision of a new clerical employee was as follows: first month, 43%; second month, 16%; third month, 9%; fourth month, 6%; fifth month, 5%. Using these and other obtained values, the following costs associated with turnover were derived: (*a*) cost of recruiting = the average cost to recruit and screen an applicant is estimated to be $20 per person; (*b*) cost of supervisory training = $593.94 (wages based on average supervisory monthly salary rate of $781.50 for 1972); (*c*) cost of loss of employee productivity during training period = $667.50 (wages based on average monthly salary rate of $445.00 for 1972, cost of loss of productivity based on 50%, 40%, 30%, 20%, and 10% of wage for first to fifth month respectively); (*d*) cost of fringe benefits during training period = $117.

In order to calculate the total cost, it was necessary to know the selection ratio and base rate. The total number of permanent full-time positions for the intermediate clerical class series was about 400. The selection ratio was found to be .24 (about 1,700 applicants applying for 400 job openings during a 25-month period). It was also found that about 50% of the clerical employees were long term, 20% intermediate term, and 30% short term.

With 400 positions to fill and a base rate of 50%, we could assume at least 200 long-term employees were needed to replace the intermediate- and short-term employees terminating during this period. Based on this assumption, the cumulative percentage of distribution of each tenure group, the total number of applicants who had to be screened and hired, the numbers hired in each tenure group, the selection ratio, the recruiting costs, the training costs, the total costs, and the average cost per long-term em-

TABLE 1

NUMBER OF APPLICANTS SCREENED AND HIRED AND RELATED COST ESTIMATES FOR DIFFERENT CUTTING SCORES ON THE WEIGHTED APPLICATION BLANK

Score	Cumulative %s scoring higher			Number screened	Number hired				Selection ratio	Total cost ($)	Average cost per long-term employee ($)
	Long term	Inter-mediate	Short term		Long term	Inter-mediate	Short term	Total			
0 (Existing procedures)				1,700	200	80	120	400	.24	585,376	2,926
5	.98	.95	.74	408	200	77	90	367	.90	514,047	2,570
8	.95	.60	.50	421	200	50	63	313	.74	439,871	2,199
10	.72	.30	.24	555	200	33	39	272	.49	386,035	1,930
12	.42	.15	.06	952	200	28	17	245	.26	356,757	1,783
14	.30	.05	.03	1,333	200	13	11	224	.17	335,430	1,677
16	.12	.05	.00	3,333	200	33	0	233	.07	387,836	1,939
17	.10	.00	.00	4,000	200	0	0	200	.05	355,687	1,778

ployee were calculated and are shown in Table 1 for each cutting score of the weighted application blank.

The data in Table 1 were computed by a revised method described by Dunnette (1966) and England (1971). The computations shown below, for a score of 12, illustrate the method used:

1. It can be seen from Table 1 that 42% long-term, 15% intermediate-term, and 6% short-term applicants would be selected at a score of 12 and above.
2. Using a cutting score of 12, the total number of applicants which would have to be screened = (Desired long-term employees)/(Base rate for long-term employees) × (% long-term employees hired) = 200/(.50) (.42) = 952.
3. The number of potential intermediate-term employees = (Number screened) × (Base rate for intermediate employees) = (952) (.20) = 190, of whom (190) × (% intermediate-term employees hired) = (190) (.15) = 28 would be hired.
4. The number of potential short-term employees = (952) (Base rate for short-term employees) = (952) (.30) = 286, of whom (286) × (% short-term employees hired) = (286) (.06) = 17 would be hired.
5. Hence, the total hired = 200 + 28 + 17 = 245.
6. The selection ratio = 245/952 = .26.
7. The estimated total cost = (Number hired) × (Training cost) + (Number screened) × (Recruiting cost) = (245) ($1,378.44) + (952) ($20) = $356,758.
8. The cost per long-term employee = $356,758/200 = $1,783.

As is shown in Table 1, the maximum utility of the weighted application blank is at a cutting score of 14. Using a cutting score of 14, the total cost of hiring 200 long-term employees was estimated to be about $335,430. Without using the weighted application blank, the total cost of the existing selection procedures was approximately $585,376 (as shown in the first row of Table 1). Thus, the estimated maximum dollar savings expected from the use of the weighted application blank is $249,946 for the [illegible]-month period. Of course, variations in supply and demand in the labor market may require upward or downward adjustment of the cutting score which will decrease the amount of dollar savings. However, it should be noted that the utility estimates of the weighted application blank are large enough to be quite significant from a practical point of view. These figures show that even in the case of a relatively low-level job with modest hiring and training costs, the dollar savings that can be expected from the introduction of a valid weighted application blank are quite large.

REFERENCES

Berkson, J. "Cost-Utility" as a measure of the efficiency of a test. *Journal of American Statistical Association,* 1947, **52**, **75**, 24[illegible]5.

Cronbach, L. J., & Gleser, G. C. *Psychological tests and personnel decisions.* Urbana: University of Illinois Press, 1965.

Dornon, J. M. *Identification of long-tenure hourly factory workers using a weighted application blank.* (Experimental Publication System Ms No. 276-2) Washington, D.C.: American Psychological Association, 1970.

Dunnette, M. D., Kirchner, W. K., Erickson, J. R., & Banas, P. A. Predicting turnover of female office employees. *Personnel Administration,* 1960, 23, 45–50.

Dunnette, M. D. *Personnel selection and placement.* Belmont, Calif.: Wadsworth, 1966.

England, G. W. *Development and use of weighted application blanks.* University of Minnesota, Industrial Relations Center, 1971.

Equal Employment Opportunity Commission. Guidelines on employee selection procedures, *Code of Federal Regulations,* Title 29, Chapter XIV, Part 1607, 1970.

Kavanagh, M. J., & York, D. Y. Biographical correlates of middle managers' performance. *Personnel Psychology,* 1972, **25**, 319–332.

Tilton, J. W. The measurement of overlapping. *Journal of Educational Psychology,* 1937, **28**, 656–662.

(Received August 23, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. [illegible], 519-521

PREDICTION OF SCUBA TRAINING PERFORMANCE[1]

ROBERT J. BIERSNER[2] AND DAVID H. RYMAN[3]

United States Navy Medical Neuropsychiatric Research Unit,
San Diego, California

In an attempt to predict successful diving performance among those undergoing scuba training, 296 trainees of the Navy School for Divers Second Class were asked to complete a demographic questionnaire, a health inventory, and an attitude survey. Multiple correlations of .45 ($p < .001$) in a validation sample of 148 subjects and .39 ($p < .001$) in a cross-validation sample of equal size were obtained between these variables and a pass–fail performance criterion. Scales dealing with mental health and training apprehension were the most significant predictors. Successful performance was also associated with higher education, having fewer older brothers, and amount of parental criticism.

The hazardous duties of military divers include inspection of underwater structures (hulls, docks, bridges, and so forth), ship salvage, and submarine rescue. Previous research has shown that experienced divers have several distinctive personality and demographic characteristics, especially trouble with authority and difficulty in relating emotionally to others (Biersner, 1973; Biersner & Cameron, 1970). Other research has found that successful underwater demolition team training success is predictable from physical fitness and perceived health measures (Biersner, Gunderson, Ryman, & Rahe, 1972), as well as from age and pay grade (Gunderson, Rahe, & Arthur, 1972). However, little research has been done to predict successful diving performance among those undergoing scuba training. It is not known, for example, whether measures of social and emotional adjustment can be used to predict such performance.

The present research is an attempt to predict performance in a hazardous military scuba training course from attitude measures, perceptions of physical and emotional health, and demographic information.

METHOD

Subjects

Subjects were 296 trainees attending the Navy School for Divers Second Class at San Diego, California. The trainees were divided into validation and cross-validation samples of 148 subjects each. Subjects were from 13 sequential diving classes which lasted five weeks each. The trainees came from all branches of the military service, with the greatest proportion from the Navy (63%). The trainees had a modal pay grade (rank) of E–3 (seaman, Army private) and mean age of 22.3 years. The average education was 12.2 years.

Training Program

Scuba training consisted of two weeks of instruction in diving medicine and physics and three weeks of practical experience involving the use and maintenance of diving equipment. The final week consisted of ocean dives to a depth of 60 feet.

Criterion Measure

The criterion was success or failure in training (pass–fail), as obtained from the official records after each class finished the course.

Predictor Measures

A demographic questionnaire, a health inventory, and an attitude survey were administered at the start of training. The demographic questionnaire asked for the following: marital status, pay grade, branch of service, education, size of hometown, number of older brothers and sisters, number of arrests and truancies before the age of 18, and frequency of parental criticism and

---

[1] This study, Report 73-54, was supported by the Bureau of Medicine and Surgery, Department of the Navy, under Research Work Unit MF51.524.002-5011 DD5G. The opinions expressed are those of the authors and are not to be construed as necessarily reflecting the official view or endorsement of the Department of the Navy.

[2] Now at the Naval Education and Training Support Command, Pensacola, Florida.

[3] Requests for reprints should be sent to David H. Ryman, Biochemical Correlates Division, U.S. Navy Medical Neuropsychiatric Research Unit, San Diego, California 92152.

TABLE 1

VALIDITIES AND BETA WEIGHTS OF PREDICTOR ITEMS AND SCALES FOR TWO SAMPLES OF SCUBA TRAINEES

Item or scale	*r* Sample 1 (n = 148)	*r* Sample 2 (n = 148)	Beta weight[a]
Education	.16*	.28**	.09
No. Older Brothers	−.27**	−.15	−.25
Criticism and Maternal Employment	.24**	.16*	.19
Visual	−.16*	−.13	−.13
General Mental	−.25**	−.20*	−.19
Concern	−.27**	−.26**	−.15
Leadership	.22**	.16*	.13

Note. For Sample 1, $R = .45$***; for Sample 2, $R = .39$***.

[a] Standardized regression coefficients of the multiple regression equation derived from the validation sample.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

praise. In addition, the following information was obtained about the parents of each subject: marital status, education, occupation, and number of hours employed. The health inventory, a condensed version of the Cornell Medical Index, contained 100 items about physical and emotional health. Subjects were given four choices in answering this questionnaire: (*a*) never, (*b*) rarely, (*c*) occasionally, and (*d*) often (scored 1 to 4, respectively). The attitude survey consisted of 25 items about training motivation, leadership, and training apprehension. Subjects were given six choices, ranging from disagree strongly to agree strongly.

Analysis

A multiple dicriminant analysis was initially performed on the validation sample but failed to distinguish among the three failure groups (voluntary, medical, and miscellaneous drops). This indicated that common factors were associated with failure in these groups. Regression coefficients were therefore computed (comparing the combined failure groups with passes) for the validation sample using demographic items, and health and attitude scales that significantly ($p <$ .05) and independently correlated with the pass–fail criterion. Regression weights computed on the validation sample were then applied to prediction of the pass–fail criterion in the second, cross-validation sample.

RESULTS

Criterion Measure

Of the diving students 69% passed, while 12% failed because of minor physical injuries and illnesses (medical drops), and 12% voluntarily withdrew (voluntary drops). The remaining failed because of major health problems, disciplinary trouble, poor swimming ability or academic difficulties (miscellaneous drops).

Predictor Measures

Four factors were subsequently identified for the health inventory. These factors and the inter-item reliabilities of each were as follows: Visual (.72), General Somatic (.84), General Mental (.91), and Anger (.75). Three factors were found among the 25 items on the attitude survey. These factors and the inter-item reliabilities of each[4] were as follows: Leadership (.82), Concern

A multiple correlation of the three predictor measures showed three demographic items (education, number of older brothers, and a combination item consisting of maternal employment and parental criticism), two attitude scales (Leadership and Concern), and two health scales (Visual and General Mental) were significantly and independently associated with the pass–fail criterion in the first sample at or above the .05 level. It was found that the amount of parental criticism had a low but significant correlation with pass–fail, and that this association was significantly moderated by maternal employment.

Table 1 lists the correlation coefficients, beta weights and significance levels of these predictor measures. On the validation sample, a multiple regression of .45 ($p < .001$) was obtained with the pass–fail criterion using these measures. Table 1 also shows that some of these measures were significantly associated with the pass–fail criterion in the cross-validation sample, resulting in a multiple regression of .39 ($p < .001$). Using these regression equations, 70% of the predictions of pass–fail were correct in the validation sample, while 69% were correct in the cross-validation sample.

DISCUSSION

The results showed that scuba training performance can be successfully predicted using a multiple regression analysis based on demographic items and health and attitude scales.

Education and birth order have been associated with performance under hazardous diving conditions (Radloff & Helmreich, 1968). The results of the present study confirmed the association of education with performance. The higher the education level of the trainee, the more likely that he would be successful. While Radloff and Helmreich (1968) found that the older-born did not make as many actual dives as did the younger-born, the results for these scuba trainees indicated that the older-born succeeded significantly more often than did the younger-born. These results indicated that factors associated with success in training may not necessarily be related to success under actual diving conditions.

---

[4] These were inter-item reliabilities which were computed with a Spearman–Brown correction.

(.76), and Confidence (.81).

It was also found that the training performance of those scuba subjects who were criticized often as youngsters was more likely to be successful if the mother was employed. This suggests that criticism from the mother or the ways in which she mediated criticism were important in adjusting to hazardous situations many years later. It is apparently not the number of documented problems such as arrests and truancies which discriminate between success and failure, but the ways in which these problems were mediated by the parents, especially the mother.

The results of the attitude scales indicated that those scuba subjects who thought they were better leaders also passed the course more often. Additionally, the attitude results substantiated the research of Stouffer et al. (1949) which showed that soldiers who are not apprehensive about being injured had the highest ratings of combat effectiveness. Scuba trainees who were low on the Concern factor, which consisted largely of items about possible physical injury during training, performed better during training. This is further supported by the results of the health inventory which found that those scuba trainees who reported fewer visual problems and emotional symptoms at the start of training were more likely to succeed. Gunderson, Rahe, and Arthur (1972) also found that health measures could be used to predict pass–fail during hazardous underwater demolition team training.

REFERENCES

Biersner, R. J. Social development of Navy divers. *Aerospace Medicine*, 1973, **44**, 761–763.

Biersner, R. J., & Cameron, B. J. Betting preferences and personality characteristics of Navy divers. *Aerospace Medicine*, 1970, **41**, 1289–1291.

Biersner, R. J., Gunderson, E. K. E., Ryman, D. H., & Rahe, R. H. Correlations of physical fitness, perceived health status, and dispensary visits with performance in stressful training. *Journal of Sports Medicine*, 1972, **12**, 107–110.

Gunderson, E. K. E., Rahe, R. H., & Arthur, R. J. Prediction of performance in stressful underwater demolition training. *Journal of Applied Psychology*, 1972, **56**, 430–432.

Radloff, R., & Helmreich, R. *Groups under stress.* New York: Appleton-Century-Crofts, 1968.

Stouffer, S. A., Suchman, E. A., De Vinney, L. C., Star, S. A., & Williams, R. M. *The American soldier: Adjustment during Army life.* N.J.: Princeton University Press, 1949.

(Received August 27, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 4, 522–523

PREDICTING ALUMNI FINANCIAL DONATION

WARREN S. BLUMENFELD[1] AND PATRICIA L. SARTAIN

Georgia State University

The purpose of the study was to predict alumni financial donation or nondonation from demographic information available at graduation. Records of 59 donors and 59 nondonors were item analyzed. Weights were assigned and cross-validated using the records of 50 held-out donors and 50 held-out nondonors. A biserial correlation of .37 ($p < .01$) was obtained.

University alumni are probably the most common target of professional fund-raising activity. However, relatively little research has been accomplished predicting financial donation (e.g., Morris, 1970). Such research would have practical application for fund-raising organizations.

There is an extensive body of literature in applied psychology regarding the successful use of weighted personal data, such as that on application blanks, in a large variety of occupational groups (for a recent review, see England, 1971). Most of these studies attempted to predict criteria such as turnover, accidents, supervisory ratings, and/or some index of production. The use of weighted personal data has also been applied successfully toward forecasting financial credit risk (e.g., Buel, 1968). As yet, however, no attempt seems to have been made to apply weighted biographical data to the problem of predicting financial donation.

The purpose of this study was to develop and cross-validate a psychometric scoring procedure (based on demographic information available at graduation) to predict alumni financial donation or nondonation.

METHOD

Subjects

The subjects were 218 students who either graduated from or last attended Georgia State University in 1963, 1964, or 1965. Of these alumni, 109 were donors to the 1968–1969 Alumni Annual Fund. A control group of 109 nondonors was developed by selecting the name following the alumni donor in the questionnaire file (year of graduation or last attendance held constant). The overall sample was composed of sixty-eight 1963 alumni (31%), fifty 1964 alumni (23%), and one hundred 1965 alumni (46%).

Independent Variables

Demographic characteristics available at time of graduation were collected from a variety of *existing* sources. The 22 characteristics investigated were sex, whether or not the spouse attended the university, age at graduation or at date of last attendance, school attended within the university, degree from the university, type of degree obtained (certificate or associate degree, bachelor's degree, master's degree, and/or doctoral degree), whether or not financial assistance was received, participation in athletics, organization membership (religious, social, honorary, and/or professional), degree from another institution (bachelor's degree, master's degree, and/or doctoral degree), major, undergraduate grade point average, graduate grade point average, and hours transferred from another institution to the university.

Data Analysis

An item analysis was conducted with the demographic information of 59 donor records and 59 nondonor records. Assigned weights as described by England (1971) were developed. The weights were then cross-validated with the records of hold-out groups of 50 donors and 50 nondonors.

RESULTS

Item analysis indicated that 7 of the 22 characteristics appeared to differentiate between the donors and nondonors. Donating alumni were most likely to be (*a*) male, (*b*) business school students, (*c*) Georgia State University graduates, (*d*) holders of a master's degree from Georgia State University, (*e*) economics majors, (*f*) with high *or* low undergraduate grade point average, and (*g*) with moderate to high graduate grade point average. Item weights were assigned, and a biserial correlation of .37 ($p < .01$) was obtained in the cross-validation.

---

[1] Requests for reprints should be sent to Warren S. Blumenfeld, Department of Management, School of Business Administration, Georgia State University, Atlanta, Georgia 30303.

DISCUSSION

The data indicated that it is possible to forecast, at a better than chance level, the financial donation behavior of one set of alumni based on available demographic data. A behavioral profile can be developed (and cross-validated) that can pre-identify those individuals most favorably predisposed toward contribution. The agency attempting to solicit funds might segment the market in terms of (*a*) types of appeals to be made to differentially disposed individuals and (*b*) amount of attention given during subsequent fund-raising activities. Other applications of the ability to forecast donation propensity may also exist.

In conclusion, the data reported suggest that the prediction of who will and who will not contribute financially is a feasible and institutionally meaningful objective for alumni relations management (and other fund-raising-oriented agencies) to pursue.

REFERENCES

Buel, W. D. Evaluating mortgage loan risk. *Journal of Applied Psychology*, 1968, **52**, 399–402.

England, G. W. *Development and use of weighted application blanks.* (Bulletin 55) Minneapolis: University of Minnesota, Industrial Relations Center, 1971.

Morris, D. Anatomy of the big donor. *Alma Mater*, 1970, **37**(5), 23–27.

(Received June 6, 1973)

Manuscripts Accepted for Publication in the *Journal of Applied Psychology*

Comparisons of Self- and Superior Ratings of Managerial Performance. Herbert G. Heneman III (Graduate School of Business, University of Wisconsin-Madison, 1155 Observatory Drive, Madison, Wisconsin 53706).

Some Personality Correlates of the Relationships Between Individuals and Organizations. H. Russell Johnston, Jr. (Babcock Graduate School of Management, Wake Forest University, 7657 Reynolda Station, Winston-Salem, North Carolina 27109).

The Harassed Decision Maker: Time Pressures, Distractions, and the Use of Evidence. Peter Wright (Department of Marketing, Graduate School of Business, Stanford University, Stanford, California 94305).

Personal Characteristics Differentiating Female Executive from Female Nonexecutive Personnel. Robert F. Morrison (Faculty of Management Studies, University of Toronto, 246 Bloor Street West, Toronto, Ontario, Canada M5S 1V4) and Maria-Luise Sebald.

University Climate Perceptions by Black Students and White Students. C. Michael Pfeifer, Jr. (Westinghouse Behavioral Safety Center, Box 948, American City Building, Columbia, Maryland 21044) and Benjamin Schneider.

Comparison of Job Evaluation Methods: A "Policy Capturing" Approach Using the Position Analysis Questionnaire. David D. Robinson (Department of Business Administration, College of Business, Utah State University, Logan, Utah 84322), Owen W. Wahlstrom, and Robert C. Mecham.

Farm and Factory in the Kibbutz: A Study in Agrico-Industrial Psychology. Dov Eden (Faculty of Social Sciences, Tel Aviv University, Tel Aviv, Israel) and Uri Leviatan.

Predicting the Evaluation of Real Stimulus Objects from Abstract Evaluation of Their Attributes: The Case of Trout Streams. Jordan J. Louviere (Department of Geography, The Florida State University, Tallahassee, Florida 32306).

Individual Differences and Reactions to Job Characteristics. John P. Wanous (Graduate School of Business Administration, New York University, 100 Trinity Place, New York, New York 10006).

Some characteristics of Inaccessible Respondents in a Telephone Survey. Charles N. Weaver (St. Mary's University, School of Administration, 2700 Cincinnati Avenue, San Antonio, Texas 78284), Sandra L. Holmes and Norval D. Glenn.

Effect of Noise on Intellectual Performance. Neil D. Weinstein (Department of Psychology, University of California, Berkeley, Berkeley, California 94720).

Organizational Commitment, Job Satisfaction, and Turnover Among Psychiatric Technicians. Lyman W. Porter (Graduate School of Administration, University of California, Irvine, Irvine, California 92664), Richard M. Steers, Richard T. Mowday, and Paul V. Boulian.

IQ Score and Mailed Questionnaire Response. Albert J. Macek (North Star Research and Development Institute, 3100 38th Avenue South, Minneapolis, Minnesota 55406) and Guy H. Miles.

Integration of Attributes in Bus Transportation: Two Modeling Approaches. Kent L. Norman (Department of Psychology, University of Alabama, Box 2968, University, Alabama 35486) and Jordan J. Louviere.

The Relation Between Foremen's Leadership Attitudes and the Skill Level of Their Work Group. Chris Ching-Yang Hsu (General Motors Corporation, 204 Boulevard Center Building, Detroit, Michigan 48202) and Richard R. Newton.

Self-Esteem, Performance, and Satisfaction: Some Tests of a Theory. Jeffrey H. Greenhaus (Department of Management Science, Stevens Institute of Technology, Hoboken, New Jersey 07030) and Irwin J. Badin.

Note on the Utility of Certainty Weights in Expectancy Theory. Jack Feldman (Department of Management, University of Florida, 103 Matherly Hall, Gainesville, Florida 32611).

(*Continued on p. 528*)

Journal of Applied Psychology
1974, Vol. 59, No. 4, 524–526

EFFECTS OF SPECTRAL DIFFERENCES IN ILLUMINATION ON FATIGUE

JAMES B. MAAS,[1] JILL K. JAYSON,[2] AND DOUGLAS A. KLEIBER

Center for Improvement of Undergraduate Education, Cornell University

An experiment was conducted to determine the effects of different spectra of environmental illumination on fatigue after a period of studying. No significant differences in the subjects' self-reported states were identified. However, objective measures revealed less perceptual fatigue and better visual acuity under lighting which closely approximated the spectral quality of natural sunlight than under traditional cool-white lighting.

The present investigation was an initial attempt to explore the relationship between spectral differences in environmental illumination and general and perceptual fatigue in humans performing mental activity. In this case, subjects studied in a room under lighting conditions which were varied over a period of evenings. Because of the dearth of prior research on which to base assumptions, and as a conscious step to reduce experimenter expectancy effects, no specific hypotheses were advanced as to the relative advantages of the different light sources tested.

METHOD

Subjects

Twelve male and 29 female undergraduate students took part in the experiment. Subjects were chosen randomly from a larger pool of introductory psychology students who had volunteered to participate in research.

Procedure

A university classroom was converted to a study room and equipped with two independent fluorescent light sources installed in the same fixtures. The first, "Vita-Lite" (V),[3] was a full-spectrum light closely approximating the spectral quality of natural sunlight. The second was a standard cool-white fluorescent light (CW), commonly used in institutional settings. The physical structures of the lamps are identical; only upon direct comparison are they seen as differing in the color or the light they emit. An analysis of energy emitted by the two lights at each frequency is presented in Figure 1.

The lamps were installed behind standard translucent covers and controlled from a room located directly behind the classroom. The illumination of the lights was stabilized and equated at 82 footcandles.

The experiment made use of repeated measures on each subject thereby controlling for variability due to individual differences. Each subject studied in the room four hours a day, on four consecutive days. Subjects spent two days under each of the experimental conditions. To control for ordering and position effects, a counterbalanced design was used that alternated the lighting conditions from CW–V–V–CW to V–CW–CW–V across subjects.

All subjects, and the people involved in actual data collection, were kept naive to the experimental conditions and the exact nature of the dependent variables being investigated. Subjects were instructed to bring to each session enough course homework to keep them occupied for four hours, and they were asked not to leave the seminar room.

The tests for fatigue were administered to each subject during the first and last 15 minutes of each four-hour session.

In attempting to measure fatigue, this experiment considered two separate phenomena: *subjective fatigue* (feelings of weariness, ennui, and tiredness, obtainable only through self-reports) and *objective fatigue* (fatigue that can be ascertained by objective measures of decrement in performance on various tasks). Pre- and posttreatment measures included a subjective questionnaire and three objective tests.

The subjective questionnaire was composed of three sections. The first section of questions was designed to ascertain the subjects' overall readiness and fitness to study; this section differed somewhat in the prestudy and poststudy questionnaires to avoid a response set to the particular items. The second section of the question-

---

[1] Requests for reprints should be sent to James B. Maas, Center for Improvement of Undergraduate Education, Cornell University, Rand Hall, Ithaca, New York 14850.

[2] Now at Georgetown Medical School.

[3] Developed by the Duro-Test Corporation, North Bergen, New Jersey.

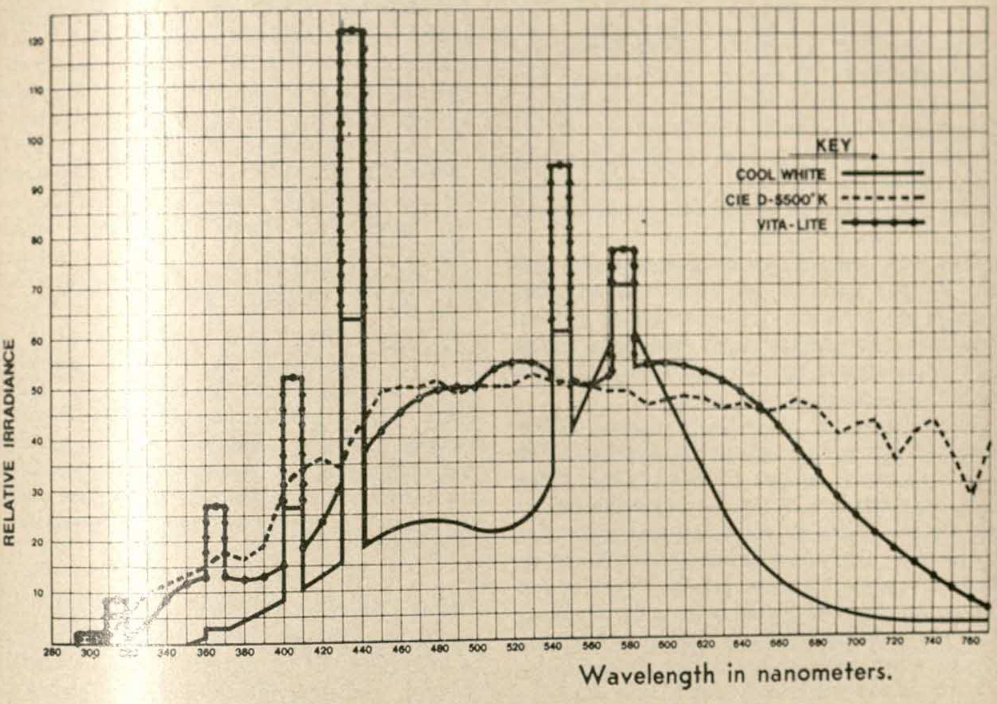


FIGURE 1. Relative spectral irradiance distributions of the fluorescent test lamps compared with natural daylight at 5500°K (as specified by the International Commission of Illumination).

naire was a checklist for measuring subjective fatigue, developed by the United States Air Force (Pearson & Byers, 1956). The checklist is composed of 13 statements concerning feelings of fatigue. Subjects were asked to check whether they felt "better than," "the same as," or "worse than" the described feeling. The final portion of the questionnaire contained 21 bipolar items relating to feelings of fatigue and general reactions to the atmosphere of the room, such as warm–cold, tense–relaxed, dark–light, and pleasant–unpleasant. A format similar to the semantic differential was used—a five-point scale with the endpoints defined by the pair of words making up each item.

The objective measures used were based on perceptual tasks, since vision is the sense most affected by study and environmental illumination. The Critical Flicker Fusion test developed by Simonson (Simonson & Enzer, 1941) was used to assess both visual and central nervous system fatigue.[4]

A Lafayette Model 12025 Critical Flicker Fusion Instrument with viewing chamber was set at 90% illumination/cycle and lamp intensity on bright. The critical flicker fusion for each subject was determined by the staircase method of limits, with the same starting point and darkened surround in the viewing chamber across both conditions.

Visual acuity was measured with a standard Snellan Eye Chart. Subjects were seated 15 feet from the chart and instructed to read as many lines as they could with one eye, then the other. The number of the line on which they made two or more errors was recorded. No feedback was given, to control for practice effects.

A pursuit rotor test was used to measure hand–eye coordination. The apparatus consisted of a Lafayette Model 30013 Photoelectric Rotary Pursuit Instrument connected to a repeat-cycle timer and a clock/counter. Subjects had to track, with a stylus, a light traveling counter-clockwise around a triangular path at 60 revolutions per minute. The test required five trials of 60 seconds each, with a 10-second rest between trials.

---

[4] For a general review of critical flicker fusion see Simonson and Brozek (1952), and Kogi and Saito (1971).

RESULTS

The first step in analysis involved examining each individual's change in performance over each four-hour study session. This was done by computing difference scores (from pretest to posttest) for all the variables of the fatigue tests. To preserve the control for order and position effects, the scores from the two sessions under the same lighting conditions were averaged together. This yielded two difference scores (CW and V) per variable for every subject. Two mean scores for each variable across all subjects were then computed to see if there were significant differences in performance measured under each of the lighting conditions.

Only one of the subjective variables (lively–lethargic) showed a statistically significant difference between lighting conditions. This difference demonstrated a tendency for subjects to become less lively or more lethargic after four hours under the cool-white light, while there appeared to be no change under the Vita-Lite condition. But the fact that it is the only difference, and is not qualitatively very different from some of the other subjective variables, suggests that this result was quite likely due to chance.

On the Photoelectric Rotary Pursuit test there were no significant differences between the two conditions. However, the other two objective measures, flicker fusion and visual acuity, showed significant differences. The presumption behind critical flicker fusion is that a neural system that can respond to high rates of intermittency with an intermittency of its own is functionally better than one that responds by continuity. Therefore, the higher the critical flicker fusion, or in the case of this experiment, the smaller the decrease over time, the better. The decrease in the critical flicker fusion was significantly less under the Vita-Lite condition than under cool-white ($p < .05$).

In measuring visual acuity, a comparison is made between what a subject can see at 20 feet and what a "normal person" can see at other distances. A high test score, for example, 100, indicates poor visual acuity, since the subject sees clearly from only 20 feet what others see from 100 feet. Difference scores for this test were computed by subtracting pretest scores from posttest scores, and the negative mean (−.65) obtained for the Vita-Lite condition suggests that subjects could see more clearly after four hours of study under the Vita-Lite while their performance under cool-white remained the same. In other words, the effect of the Vita-Lite condition was actually to enhance visual acuity over time rather than merely to minimize fatigue. This may have been due simply to adaptation, with the visual process adjusting to the conditions after a few hours of study. If so, the results suggest that it was easier to adapt to the Vita-Lite or, alternatively, that the adaptation to the cool-white light was negated by fatigue. There were no initial differences in visual acuity between the two conditions on the pretests, thus discounting the possibility that the adaptation to the Vita-Lite was related to poor pretest performance in the unusual lighting environment.

DISCUSSION

It is clear that this line of experimentation should be expanded; other light spectra should be experimentally manipulated; tasks should be changed, and individual differences should be considered. The results of this study point to the promise of further research on the visual environment and human behavior.

REFERENCES

Kogi, K., & Saito, S. A factor-analytic study of phase discrimination in mental fatigue. In K. Hashimoto, K. Kogi, & E. Grandjean (Eds.), *Methodology in human fatigue assessment.* London: Taylor & Francis, 1971.

Pearson, R., & Byers, G. *The development and validation of a checklist for measuring subjective fatigue.* (Tech. Rep. 56-115) San Antonio, Texas: U.S. Air Force Base, Randolph Air Force Base, School of Aviation Medicine, 1956.

Simonson, E., & Brozek, J. Flicker fusion frequency: Background and application. *Physics Review,* 1952, 32, 349–378.

Simonson, E., & Enzer, N. Measurement of fusion frequency of flicker as a test of fatigue of the central nervous system. *Journal of Industrial Hygiene and Toxicology,* 1941, 23, 83–89.

(Received June 18, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 4, 527–528

EFFECT OF AMBIENT ILLUMINATION ON NOISE LEVEL OF GROUPS

MARK SANDERS,[1] JERRY GUSTANSKI, AND MIKE LAWTON

California State University, Northridge

The ambient illumination level was varied in a natural setting. Unobtrusive measures of the noise produced by groups of people gathering in the area revealed a significant ($p < .001$) reduction in noise under the low-illumination condition. Suggestions for future research are presented.

Isaac, in a series of studies (Alexander & Isaac, 1965; Isaac & DeVito, 1958; Isaac & Reed, 1961), found that ambient illumination was an important determinant of motor activity in animals. The specific effect of increased illumination appears to be dependent on whether the animal is diurnal or nocturnal. Illumination increased the activity of the diurnal rhesus monkey (Alexander & Isaac, 1965; Isaac & DeVito, 1958) and decreased the activity of the nocturnal cat (Isaac & Reed, 1961).

In the present study, it was proposed that the noise level produced by people (a possible measure of arousal) could be varied by altering the level of ambient illumination.

METHOD

The experiment was carried out unobtrusively in the halls of a classroom building at California State University, Northridge. Before the dismissal of classes, students, waiting for the next class, regularly gather in the halls. They typically sit against the walls in pairs or small groups and talk. One experimenter sat along the way approximately half way down the 10 × 135 foot hallway with a General Radio 1551-C sound pressure meter concealed in a briefcase such that only he could see the instrument. The meter was set on the "C" weighting with the damper circuit engaged to eliminate the effects of peak sounds. A second experimenter unobtrusively counted the number of people in the hallway.

Prior to creation of the illumination conditions, and before students began to gather, background noise levels were obtained. Only two illumination conditions could be created due to electrical wiring constraints. The high-illumination condition (the natural environmental condition) consisted of having all 20 large overhead fluorescent light panels (48 inches × 24 inches) in the hallway lit. This produced from 15 to 25 footcandles of illumination measured at waist height by a Weston Illumination Meter, Model 756 with Viscor Filter. The low-illumination condition was created by turning off two thirds of the light panels in an on-off-off pattern. This provided an irregular pattern of illumination. At waist height, the illumination under the "off" panels was 1 footcandle, under the "on" panels the illumination was 25 footcandles. Halfway between the on and off panels the illumination measured 8 footcandles indicating a rapid drop-off in illumination as distance from the on panels increased.

The lights were turned off for the low-illumination condition 45 minutes before the observation began, and the light switch was taped with a note marked "loose wire." This was done to prevent students or faculty from turning on the lights during the experimental condition.

Under the high- and low-illumination conditions 10 observations were made. Time of day, day of week, and location were equated in the two conditions.

RESULTS

Even though a damper circuit was used on the sound pressure meter, readings still varied over a 2 to 5 decibels (C scale) range. The midpoint was therefore used as the data point. Decibel readings in both the background and illumination conditions were converted to sound pressure ratios. Background sound pressure ratio was then subtracted from the sound pressure ratio recorded under the appropriate illumination condition, and the difference was converted back to decibels (Peterson & Gross, 1967).

The mean noise level in the low-illumination condition was 50.3 decibel (C scale) with an average of 12 people present. The mean noise level in the high-illumination condition was 61.1 decibel (C scale) with an average of 12 people present. An analysis of covariance was performed on the data, using the number of people present as the covariate. The difference in mean noise levels was significant at the .001 level ($F = 33.1$, $df = 1/17$).

---

[1] Requests for reprints should be sent to Mark S. Sanders, Department of Psychology, California State University, Northridge, California 91324.

DISCUSSION

The results indicate that ambient illumination can effect the noise generated by people in the environment. Due to the unobtrusive nature of the experiment, several important variables were not controlled. First, the duration of exposure to the altered environment (i.e., low illumination) could not be controlled. Some students entered the darkened environment immediately before measurements were made, while others had been in the hallway for up to 30 minutes.

Further research should investigate the effects of the duration of exposure. It is possible that the reduced noise level in the low-illumination condition may be a transitory phenomenon. Possibly, with prolonged exposure, noise levels in high- and low-illumination environments may become more nearly equal. Nevertheless, even if the effect is temporary, there may still be some practical utility in reducing ambient illumination in areas, such as hallways, where it is desirable to reduce the noise generated by people in the area.

In the present study, the low-illumination condition resulted in a heterogeneous illumination environment characterized by "hot" and "cold" spots. The high-illumination condition, on the other hand, did not possess such a varied environment. In essence then, level of illumination was confounded with variability of illumination. Further research should be aimed at unconfounding these variables by constructing an environment in which illumination level can be varied without disturbing the pattern of illumination.

It is felt that results from this and future research could be applicable to building design and to the growing field of architectural psychology.

REFERENCES

Alexander, M., & Isaac, W. Effect of illumination and d-amphetamine on the activity of the Rhesus macaque. *Psychological Reports*, 1965, **16**, 311-313.

Isaac, W., & DeVito, J. Effect of sensory stimulation on the activity of normal and pre-frontal lobectomized monkeys. *Journal of Comparative and Physiological Psychology*, 1958, **51**, 172-174.

Isaac, W., & Reed, W. The effect of sensory stimulation on the activity of cats. *Journal of Comparative and Physiological Psychology*, 1961, **54**, 677–678.

Peterson, A., & Gross, E. *Handbook of noise measurement* (6th ed.) West Concord, Mass.: General Radio Co., 1967.

(Received July 16, 1973)

(*Continued from page 523*)

Effects of Readability of Context Passages on Subsequent Inspection Rates. Ernst Z. Rothkopf (Learning and Instruction Research Department, Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974) and Richard P. Coatney.

Organizational Climate as a Function of Personal and Organizational Variables. James F. Gavin (Department of Psychology, Colorado State University, Fort Collins, Colorado 80521).

The Effects on Response Rates of Formal and Informal Questionnaire Follow-Up Techniques. John F. Anderson (University Measurement Services Center, University of Minnesota, 9 Clarence Avenue S.E., Minneapolis, Minnesota 55414) and Douglas R. Berdie.

"Practical Significance" of Locke's Theory of Goal Setting. Gary P. Latham (Human Resources Research, Weyerhaeuser Company, Tacoma, Washington 98401) and J. James Baldes.

Using Subordinate Ratings to Elicit Behavioral Changes in Supervisors. W. Harvey Hegarty (Graduate School of Business, Indiana University, School of Business Building, Bloomington, Indiana 47401).

Work Behavior Dimensions of Psychiatric Attendants and Aides. M. K. Distefano, Jr. (Central Louisiana State Hospital, Division of Mental Health, P.O. Box 31, Pineville, Louisiana 71360) and Margaret W. Pryer.

Behavior Modification and Absenteeism: Intervention in One Industrial Setting. Victor U. Gamboa (Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, Michigan 48106) and Ed Pedalino.

Effect of Rate and Level of Lowered Finger Surface Temperature on Manual Performance. John M. Lockhart (Research Psychologist, Pioneer Research Laboratory, US Army Natick Laboratories, Natick, Massachusetts 01760), Harold O. Kiess, and Thomas J. Clegg.

An Analysis of Situational Moderating Effects on the Relationships Between LPC and Other Psychological Measures. Marshall Sashkin (Department of Management, School of Administration, Wayne State University, Detroit, Michigan 48202), F. Carter Taylor, and Rama C. Tripathi.

Formats for Presenting Procedural Instructions. Arthur S. Blaiwes (Department of the Navy, Naval Training Equipment Center, Orlando, Florida 32813).

The Protestant Ethic Scale as a Predictor of Repetitive Work Performance. Matthew R. Merrens (Department of Psychology, State University, Plattsburgh, New York 12901) and James B. Garrett.

Interaction of Subordinate and Leader Characteristics in Moderating the Consideration–Satisfaction Relationship. David M. Herold (College of Industrial Management, Georgia Institute of Technology, Atlanta, Georgia 30332).

The Relationship Between Consumers' Category Width and Trial of New Products: A Reappraisal. Lyman E. Ostlund (Marketing Department, Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19174) and Brynulf Tellefsen.

Journal of Applied Psychology
1974, Vol. 59, No. 4, 529–531

VISUAL DEPENDENCY IN THE ERECT AND SUPINE POSITIONS

JACOB H. LICHTENSTEIN[1] AND RAYFORD T. SAUCER[2]

NASA Langley Research Center, Hampton, Virginia

The field dependency-independency of 37 subjects was examined under supine and erect body position. Female subjects were more field dependent than were males. No significant differences were found between pilot and nonpilot males. Male subjects exhibited greater field dependency in the supine position than in the erect position. The results suggest the need to consider three separate measures of field dependency (slope, breakpoint, and variability). Possible implications for the use of subjects in simulators involving visual and body motion cues are discussed.

A number of simulators and devices emphasize a dependence of the subject on visual or bodily cues. It seems useful to know (*a*) if there are individual differences in the degree to which subjects on these devices depend on such cues; and (*b*) what conditions of simulator use affect these differences.

The work of Asch and Witkin (1948) provides a method for measuring the visual field dependence-independence of individuals. Some encouragement that this characteristic may be relevant to pilot performance comes from work by Cullen, Harper, and Kidera (1969) who found that airline pilots generally were less field dependent than were engineers, and by Benfari and Vitali (1965) who found that field independent pilots performed a tracking task better, and that field dependent pilots were unduly influenced by the frame around the visual scene.

The objectives of the present study were to obtain preliminary data on the field dependency of pilots and personnel, at the Langley Research Center, Hampton, Virginia, used as simulator subjects. Specifically, the distribution and degree of field dependence were compared for male, female, and pilot groups, and for two task conditions frequently examined in simulators, namely, supine versus erect position of the subject. An additional aim was the examination of alternate measures of field dependency.

METHOD

Subjects

A total of 37 subjects, both male and female, ranging in ages from 22 years to 52 years, was tested. The group consisted of 8 females and 29 males; 9 of the group were male pilots.

Procedure

For the field dependency-independency test, a version of Witkins's rod-and-frame apparatus was used. The tests were conducted with the subject in two positions: (*a*) erect, where both gravity and body cues were available; and (*b*) supine, where only the cue of body position was available for aid in alining the rod. In the erect condition, the platform was horizontal and the subject sat erect, while in the supine condition, the platform was tilted up to 90 degrees, and the subject was on his back. The erect position tests were run first as they were considered less disturbing to the subject.

The equipment differed from Witkins's basic setup in that the rod and frame were electrically driven and could be set at any desired angle between −45 degrees and +45 degrees by remote control, with the resulting angles read from remote meters. The chair for the subject and the rod-and-frame apparatus were mounted at opposite ends of a platform. This permitted tilting the entire platform, chair, and apparatus up 90 degrees while maintaining the same relationship between the subject and the rod and frame in both the supine and erect positions.

The subject was placed in either the erect or supine position. He was allowed about three minutes to become acclimated to the dark, during which time he was informed that the rod and frame would be set at various angles to the vertical (sometimes at the same angle and sometimes at different angles). The subject was to set the rod back to the vertical using a switch. For the supine tests, he was told to return the rod to a position that would be equivalent to the vertical in the erect position. The subject also was requested to close his eyes during the time

---

[1] Requests for reprints should be sent to Jacob H. Lichtenstein, Aerospace Dynamics Branch, Flight Dynamics and Control Division, NASA Langley Research Center, Hampton, Virginia 23665.

[2] Now at the U.S. Veterans Administration, Veterans Administration Hospital, Perry Point, Maryland.

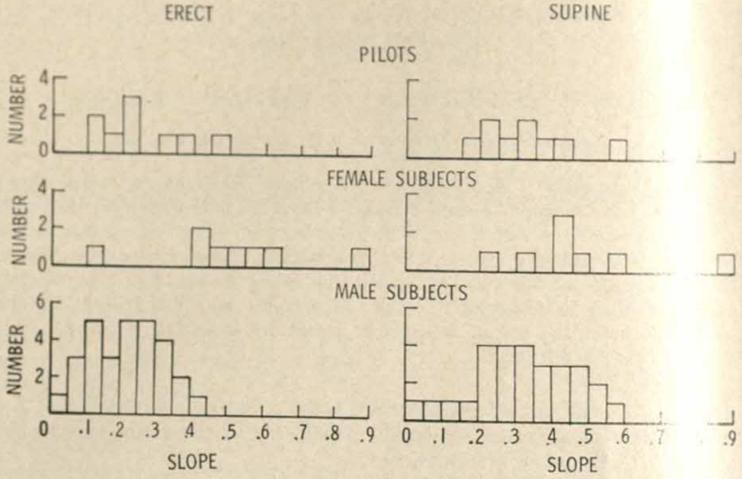


FIGURE 1. Histograms of the slope distributions for the test groups.

the rod and frame were being set for each new pair of test angles.

The range of angles used for these tests was ±40 degrees for the rod and ±45 degrees for the frame. The same 55 different rod-and-frame angle combinations were presented to each subject, although a different order of presentation was used in each case. Auditory cues for setting the angles were minimized. These tests usually ran between 45 and 60 minutes.

RESULTS AND DISCUSSION[3]

An inspection of all of the graphical presentations of subject response showed that the scores, rather than dividing into two groups, field independent and field dependent, were distributed across the range from those weakly influenced by the visual field to those strongly influenced. It was also found that for most subjects, through the central angle region, the frame exerted a nearly linear effect on the subjects' responses up to a breakpoint between 15 and 25 degrees.

Theoretically, a completely field dependent person's response in locating the rod is identical with the angle of the frame (slope = 1), whereas a completely field independent person's rod response is the same no matter what angle the frame was (slope = 0). The slope, therefore, is an indicator of the influence that the field exerts on the subject; the breakpoints, on the other hand, indicate how far this influence will extend. This suggests that a criterion for describing a person's dependency on the visual field should include both the slope and breakpoint information and perhaps the standard deviation. A choice of response at a single angle, particularly if it is beyond the breakpoint value, could lead to an entirely different conclusion from the choice of another angle only a couple of degrees away.

The slope measure developed is the slope of the best fit straight line through the data for the response of the rod against the tilt of the frame, obtained by an unweighted least squares procedure.

Figure 1 presents histograms of the slope data obtained for subjects in both the erect and supine positions for male, female, and pilot subjects. Figure 2 presents histograms of the breakpoints for these same groups.

The data showed that the scores for the entire test sample are disproportionally distributed toward field independency. This is reflected by the median slope values falling at .25 (erect) and .35 (supine), rather than .50.

In the erect position, the slopes for the eight females tested were generally among the highest of the population. Although it was originally surmised that the pilots would be less field dependent than the general group, the slope distribution for the nine pilots was similar to that for all the male subjects.

The comparison of subjects in the supine position (with the consequent reduced gravity cues) with subjects in the erect position showed that for the male subjects, pilots included, the supine

---

[3] The full set of data obtained from the tests for all subjects in both the erect and supine positions is presented in a report by Lichtenstein and Saucer (1972), which is available from National Technical Information Service, Springfield, Virginia 22151, as NASA TN D-6883 at $3.00.

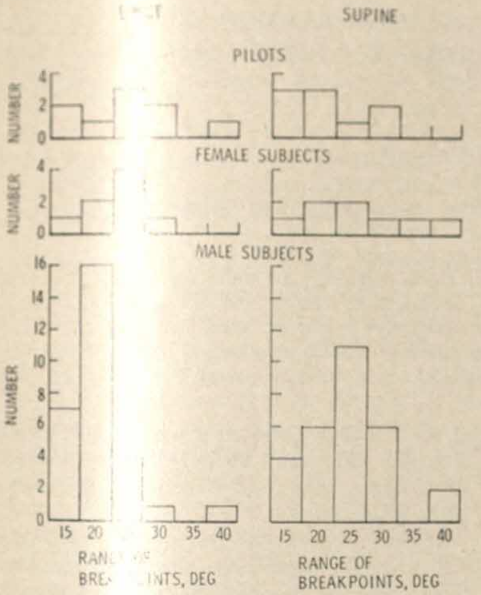


FIGURE 2. Histograms of the breakpoint distributions for the test groups.

position increased dependence on the visual field. The increase in slope was significant beyond the .001 level for the entire group and beyond the .025 level for the pilots. The average slope for the group increased from .21 to .34. On the other hand, for female subjects changes in body position showed no effect (slopes of .48 and .49, respectively).

The method of evaluating field dependency for the limited angular displacement by using only the slope data was compared with the generally accepted method for determining field dependency (Asch et al., 1948). The relative ranking of the subjects from field dependent to independent, using the two methods showed little difference for subjects at the extremes of the field dependence-independence distribution. The rank difference coefficient of correlation was computed for the upper and lower third of the ranking, and the average value was $r = .81$. However, for the subjects between the extremes, the correlation was low ($r = .20$).

In the supine position the relative ranking between the two methods exhibited the same pattern, better agreement at the extremes and poorer in the center but at a much lower level of correlation. In the present context the correlation coefficients for the extremes and center were $r = .59$ and $.03$, respectively.

This pattern of agreement between the two methods is understandable because for the field dependent subjects the high slope results in large angles and for the field independent subjects both the slope and angles are low. However, for the subjects in the center of the distribution both the slope and breakpoint are important, and using only one type of criterion can be misleading. In view of these data, it appears that a method for determining the field dependency or independency should include the slope, breakpoint, and variation data.

Possible application of the field dependency information to simulator operation can be divided into two areas. If the simulator has a rather limited motion, say ±15 degrees or less, the essential field dependency parameter that is applicable would appear to be the slope. The breakpoints, usually occurring at larger angles, probably have only a minor influence. If on the other hand, the simulator has large angular displacements greater than ±15 degrees, then both the slope and breakpoint should be considered.

The question of whether field dependent or field independent people actually perform differently on simulators or vehicles involving visual and bodily cues still needs to be determined. The present study at least suggests some factors which may need to be taken into account.

REFERENCES

ASCH, S. E., & WITKIN, H. A. Studies in space orientation. I. Perception of the upright with displaced visual fields. *Journal of Experimental Psychology*, 1948, **38**, 325–337.

BENFARI, R., & VITALI, P. The relationship between vertical orientation in the rod and frame test and vertical orientation in a compensatory tracking task. Bethpage, N.Y.: Gruman Aircraft Engineering Corporation, Research Department Memorandum RM-260J, 1965.

CULLEN, J. F., HARPER, C. R., & KIDERA, C. J. Perceptual style differences between airline pilots and engineers. *Aerospace Medicine*, 1969, **40**(4), 407–408.

LICHTENSTEIN, J. H., & SAUCER, R. T. Experimental investigation of the visual field dependency in the erect and supine positions. (NASA TN D-6883) Springfield, Va.: National Technical Information Service, 1972.

(Received May 1, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 4, 532–534

EFFECTS OF NUMERICAL DATA UPON READING TIME AND RECALL OF SCIENTIFIC WRITING

WILLIAM WITT[1]

University of Wisconsin

Traditional "readibility" studies investigating textual difficulty as an important barrier in the communication of scientific information have focused almost exclusively upon the word unit. This study examined the effects of numerical units upon reading time and recall. Half of the subjects, 100 university students, were exposed to text containing numerical references, and the other half were exposed to identical text with descriptive modifiers substituted for the numbers. Numerical data significantly increased reading time but, although in the predicted direction, fell short of significance in decreasing recall.

Textual difficulty is an important consideration in the translation of science and environmental information for public consumption. It is especially so in view of the importance of an informed citizenry in a technological society. Factors which increase difficulty can be expected to decrease comprehension and to weaken the already tenuous citizen role in support and decision making in this area.

Funkhouser and Maccoby (1971) suggested positive but indirect relationships between readability, enjoyment, and information gain in a study of science articles. Relationships between message content and degree of comprehension have been studied in terms of many factors, and there is little doubt that difficulty inhibits comprehension (Weiss, 1969).

The examination of numerical data as a factor influencing textual difficulty, however, has been neglected. The few studies that have been done are correlational, and they show extremely tentative results. Donohue, Tichenor, and Olien (1970) found indications of possible curvilinearity in the relationship between numerical units and understanding of responses to news stories. Thus, numerical data may enhance understanding up to a point beyond which understanding declines.

It is suggested that numerical data should increase difficulty by raising the level of precision by which textual substantives are modified. It simply becomes a case of having to decode more information in a given message. For example, there is less uncertainty and therefore more information in the sentence, "There are five to six quadrillion tons of air," than there is in the sentence, "There are millions upon millions of tons of air," though they are equivalent in number of words and almost so in syllables (Anderson, 1970; Bross, Shapiro, & Anderson, 1972; Flesch, 1946; Klare, 1963).

Reading material which contains more information should be harder to remember and should take longer to decode. The two indicators of difficulty, reading time and recall, therefore should show significant differences with numerical data present and absent. The present study examined the effects of numerical data on reading time and recall in a popularized science article. Specifically, it was hypothesized that with numerical data present (*a*) reading time should increase, and (*b*) recall should decrease when compared to equivalent text without numerical data. Unlike previous studies, the conventional measure of difficulty, readability (Klare, 1963), was controlled. This allowed only numerical data to vary in the article.

METHOD

Two versions of an article about air pollution were prepared from a published text (Malone, 1970). They were identical except that one carried 12 numerical references and the other carried descriptive modifiers where numbers occurred in the first. All numerical references were in familiar units, and no complex mathematics were involved.

For example, the version without numerical data reported, "Man is sensitively dependent upon that part of the air composed of oxygen, and cannot survive much more than a few minutes without it." The version with numerical data reported, "Man is sensitively dependent upon that one-fifth of the air composed of oxygen, and cannot survive much more than six minutes without it."

Otherwise, they were identical in text, length (200 words), number of sentences (10), and practically so in readability as measured by the Flesch score: without numbers = 6.3374, with numbers = 6.9161 (Powers, Sumner, & Kearl, 1958).

---

[1] Requests for reprints should be sent to William Witt, Department of Communication, University of Wisconsin, Stevens Point, Wisconsin 54481.

The experiment was of posttest-only design in which respondents answered a brief demographic questionnaire, read one article, and completed a short-answer 10-question quiz on the facts presented in the article. Respondents were 100 university students, half of them male, all members of introductory communications classes at the University of Wisconsin—Stevens Point. Testing was conducted during the academic session of 1971–1972.

Each respondent received all portions of the experiment in a packet and, upon signal, proceeded from beginning to end at his own pace. Each recorded the time he began and finished reading the article. All packets contained identical material except for the article. Half of the respondents received the article with numerical data, and the other half received the article without.

RESULTS

Results were coded and computer analyzed (Nelson, 1971). The order in which units occurred in the packet had been varied systematically and was found to be nonsignificant. Sex was also found to be nonsignificant.

General interest in the subject area of the article, conservation, was measured in the demographic questionnaire on a 5-point (0–4) scale which had been developed and validated earlier (Witt, 1968).

Numerical data were found to have a significant effect, in the predicted direction, upon article reading time, with recall and interest controlled ($5.71 \leq F \leq 7.15$). However, the effect upon recall, though in the predicted direction, failed to reach significance, with time and interest controlled ($.59 \leq F \leq 3.55$). In fact, interest almost reached significance ($F = 3.87$), and numerical data almost went to zero with article-reading time uncontrolled. Interest showed no significant effect upon reading time (Table 1).

DISCUSSION

While results indicated that numerical data did increase reading time, the increase appeared primarily among individuals of lower rather than higher recall (low: 53.4–66.6 seconds versus high: 58.4–58.8 seconds—see Table 1). If we can view on a post hoc basis those who exhibited higher recall as individuals with greater potential for comprehension, it would appear that numerical data is a barrier with respect to reading ease primarily for the lower potential individuals. This appears to coincide rather well with the restricted use of numerical data in science writing in the mass media, which aim at the "lowest common denominator" of citizen level, and in the more generous use of numerical data in specialized publications as, for example, the *Scientific American*, which aim at a higher potential audience.

Numerical data may indeed lower recall, but with article reading time controlled statistically rather than experimentally (respondents' reading time was not limited in this study) results failed to reach significance. Control of time of exposure on an experimental basis might be desirable for more positive results. Interest in the subject area also might respond favorably as an independent variable under these conditions, since it would seem that heightened interest might well enhance recall.

TABLE 1

EFFECTS OF NUMERICAL DATA UPON READING TIME AND RECALL UNDER SEVERAL CONTROL CONDITIONS

Control	Numerical data		F			
	Absent	Present	Column	Row	Column × Row	df
Reading time^a						
Recall^b						
Low (0–7) High (8–10)	53.4 58.4	66.6 58.8	7.15**	1.19	1.47	1/99
Interest^c						
Low (0–1) High (2–4)	55.8 57.0	60.3 60.5	5.71*	.24	.00	1/100
Recall^b						
Time^a						
Low (30–60) High (75–105)	7.05 7.54	6.54 6.72	3.55^d	.95	.19	1/100
Interest^c						
Low (0–1) High (2–4)	6.60 7.52	6.47 7.05	.59	3.87^d	.13	1/100

^a Time taken to read air pollution article, in seconds.

^b Mean No. items correctly recalled in 10-item questionnaire.

^c Scaled from 0 to 4 on generalized scale of interest and activity in conservation developed by the author (Witt, 1968).

^d If $F = 3.94$, $df = 1/100$, then $p < .05$.

* $p < .05$.

** $p < .01$.

REFERENCES

Anderson, R. L. Rhetoric and science journalism. *Quarterly Journal of Speech*, 1970, **56**, 358–368.

Bross, I. D., Shapiro, P. A., & Anderson, B. B. How information is carried in scientific sub-languages. *Science*, 1972, **176**, 1303–1307.

Donohue, G. A., Tichenor, P. J., & Olien, C. N. Attitudes and mass media learning about two environmental issues. Paper presented at Symposium on Mass Communication Research and Environmental Decision-Making, Madison, Wisconsin, January 22–23, 1970.

Flesch, R. *The art of plain talk*. New York: Harper & Row, 1946.

Funkhouser, G. R., & Maccoby, N. Communicating specialized science information to a lay audience. *Journal of Communication*, 1971, **21**, 58–71.

Klare, G. R. *The measurement of readability*. Ames: Iowa State University Press, 1963.

Malone, T. F. The atmosphere. In H. D. Johnson (Ed.), *No deposit–no return, man and his environment: A view toward survival.* Reading, Mass.: Addison-Wesley, 1970.

Nelson, K. *Between-subject analysis of variance.* (FORTRAN IV program for IBM 1130, SABCA) Academic Computing Center, University of Wisconsin—Stevens Point, December 1971.

Powers, R. D., Sumner, W. A., & Kearl, B. E. A recalculation of four readability formulas. *Journal of Educational Psychology*, 1958, **49**, 99–105.

Weiss, W. Effects of the mass media of communication. In *Handbook of social psychology*. (2nd ed.) Vol. *5*. Reading, Mass.: Addison-Wesley, 1969.

Witt, W. Variables in conservation communication. Unpublished manuscript, University of Wisconsin—Madison, 1968.

(Received April 16, 1973)

ERRATUM

In the article by Michael R. Cooper, Charles E. Boltwood, and Robert J. Wherry, Sr. entitled "A Factor Analysis of Air Passenger Reactions to Skyjacking and Airport Security Measures as Related to Personal Characteristics and Alternatives to Flying," which appeared in the June 1974 issue (Volume 59) of the *Journal of Applied Psychology,* Footnote 2 on page 365 appeared incorrectly. It should read: The authors wish to thank Victoria Fein Bernstein for her invaluable assistance during the preliminary phases of the study, Thomas M. Ostrom for his comments on the construction of the questionnaire, and Milton D. Hakel, Michael T. Wood, and the late S. Rains Wallace for their constructive criticisms of earlier drafts of this report.

Journal of Applied Psychology
1974, Vol. 59, No. 4, 535–536

TYPES OF TRANSMITTAL LETTER AND QUESTIONNAIRE COLOR AS TWO VARIABLES INFLUENCING RESPONSE RATES IN A MAIL SURVEY

MICHAEL T. MATTESON[1]

University of Houston

The effects on response rates of using a form versus a semipersonal transmittal letter and a white versus a colored questionnaire were investigated. Questionnaires were mailed to 2,123 members of a national professional organization. A significantly greater number ($p < .001$) of returns were received from the groups receiving the semipersonal letter. With respect to color of questionnaire, no significant differences in rate of return were found among the groups receiving the semipersonal letter, while among the groups receiving the form letter a significantly greater number ($p < .05$) of colored questionnaires were returned.

A number of research texts contain discussion of and summarize research related to variables influencing response rates in mailed questionnaires (e.g., Miller, 1970; Parten, 1950; Simon, 1969). Such discussions generally include suggestions for increasing the rate of return by use of follow-up mailings, inducements, timing of questionnaire arrival, proper questionnaire format and appearance, and type of return envelope provided, to name but a few.

The present study focuses on differences in response rates as a function of having received (*a*) a form versus a semipersonal transmittal letter and (*b*) a colored versus a white questionnaire. It was hypothesized that (*a*) a significantly greater percentage of those receiving a semipersonal transmittal letter would respond than those receiving a form letter and (*b*) among those groups who received the same type of transmittal letter, a significantly greater percentage of those receiving the colored questionnaire would respond than would those receiving the white questionnaire.

METHOD

Subjects

The subjects were 2,123 members of a national professional organization. This sample represented 90% of the total membership of the organization and 100% of the membership residing within the continental United States, as determined by the organization's most recent membership roster. The organization is comprised primarily of individuals holding advanced academic degrees, approximately 95% of whom hold teaching, research, and/or administrative positions in colleges and universities.

Procedures

A two-page (legal size) open-ended questionnaire was developed that requested the respondent to list title, author, and publishing information for up to eight examples of what in his opinion represented significant contributions to the literature in his field. The time needed to complete the questionnaire could vary from approximately 10 minutes to an hour or more, depending on how many examples were listed and how long the respondent pondered his choices. The questionnaires were printed by an offset process on good quality paper. Half of the questionnaires were printed on white paper, the other half on pink. Pink was chosen because, of the colors available, it was thought to have the highest intensity (brightness).

The transmittal letter was on letterhead stationery, also printed offset. It contained a brief discussion of the questionnaire, the uses to which the collected data might be put, and a promise to send a summary of the results to those respondents requesting it. Half of the letters were printed with a "Dear Colleague" salutation and a facsimile signature at the closing. The remaining letters had the addressee's address individually typed at the top of the letter, contained an individually typed personalized salutation and were individually signed by the investigator. All other details in the two sets of letters were identical.

The subjects were randomly assigned to four groups. Group 1A received the semipersonalized letter and a white questionnaire; Group 1B, the semipersonalized letter and pink questionnaire;

---

[1] Requests for reprints should be sent to Michael T. Matteson, College of Business Administration, University of Houston, Cullen Boulevard, Houston, Texas 77004.

TABLE 1

PERCENTAGE OF QUESTIONNAIRES RETURNED

Transmittal letter	White questionnaire	Colored questionnaire	Total
Semipersonal	31.5	32.3	31.9
Form	19.7	24.3	22.0

Group 2A, the form letter and white questionnaire; Group 2B, the form letter and pink questionnaire. Group 2B consisted of 530 subjects with the remaining groups each containing 531. A postage-paid business reply envelope was included in each mailing. Two slightly different mailing addresses were used on the reply envelopes so that responses could be identified as coming from a Group 1 or 2 respondent. All 2,123 questionnaires were mailed bulk rate.

RESULTS AND DISCUSSION

Eighty-three of the mailings were returned by the post office as undeliverable. This reduced the sample size to 508 in Group 1A, 510 in 1B, 512 in 2A, and 510 in 2B. Of the remaining 2,040 mailings assumed to have reached their intended recipients, 550 were completed and returned for an overall response rate of 27.0%.

Table 1 summarizes the results. Among those having received the semipersonal letter, 31.9% (325/1,018) responded as compared to a 22.0% (225/1,022) response from the form letter groups. This difference, significant beyond the .001 level ($t = 5.06$), supports the first hypothesis. Such a difference indicates practical significance as well in that the slightly greater expense involved in "personalizing" one part of the mailing appears to have led to an increase in returns of better than 44% (100/225) over that of the form letter group.

The second hypothesis, that among those having received the same type of transmittal letter a higher percentage of those receiving the colored questionnaire would respond than would those receiving the white questionnaire, was supported only for the group receiving the form letter ($t = 1.77$, $p < .05$). Among this group, those receiving the colored questionnaire showed an increase in returns of almost 23% (23/101) over those having received the white questionnaire. Thus the eye-catching effect of color may operate to slightly improve returns where response rates are low, such as was the case with the form letter group. The fact that color made no apparent difference in the response rate among those receiving the semipersonal letter suggests that the influence value of the personalization, the effect of which is perhaps to increase the recipient's personal identification with the survey project, supersedes the effect of color as a more generalized, and thus nonpersonal, eye-catching device.

In summary, these data lend support to the importance of personalizing the transmittal letter in a mail survey and suggest further that in situations where a form letter must be used, with its resulting lower rate of return, a colored questionnaire may improve the response rate significantly but not to the level reached by the use of the more personalized format.

REFERENCES

Miller, D. C. *Handbook of research design and social measurement.* (2nd ed.) New York: David McKay, 1970.

Parten, M. *Surveys, polls, and samples: Practical procedures.* New York: Harper, 1950.

Simon, J. L. *Basic research methods in social science: The art of empirical investigation.* New York: Random House, 1969.

(Received June 1, 1973)

JOURNAL OF APPLIED PSYCHOLOGY

Copyright © 1974 by the American Psychological Association, Inc.

October 1974

Vol. 59, No. 5

849
27/1./75
Lib
Bureau Ed Research

Notice to APA Authors 537

ARTICLES

Effects of Intermittent, Moderate Intensity Noise Stress on Human Performance *George C. Theologus, George R. Wheaton, and Edwin A. Fleishman* 539

Effect of Noise on Intellectual Performance *Neil D. Weinstein* 548

The Harassed Decision Maker: Time Pressures, Distractions, and the Use of Evidence *Peter Wright* 555

Expert Judgment: Some Necessary Conditions and an Example *Hillel J. Einhorn* 562

Predicting the Evaluation of Real Stimulus Objects from Abstract Evaluation of Their Attributes: The Case of Trout Streams *Jordan J. Louviere* 572

Statistical Procedures for Evaluating the Practical Utility of a Moderator Approach to Prediction *Alan L. Gross, Jane Faggen-Steckler, and Karen McCarthy* 578

Utilization of Manpower: Development and Evaluation of an Assessment-Classification Model for Matching Individuals with Jobs *Lyle F. Schoenfeldt* 583

Farm and Factory in the Kibbutz: A Study in Agrico-Industrial Psychology *Dov Eden and Uri Leviatan* 596

Organizational Commitment, Job Satisfaction, and Turnover Among Psychiatric Technicians *Lyman W. Porter, Richard M. Steers, Richard T. Mowday, and Paul V. Boulian* 603

Predicting Absenteeism and Turnover: A Field Comparison of Fishbein's Model and Traditional Job Attitude Measures *John E. Newman* 610

Individual Differences and Reactions to Job Characteristics *John P. Wanous* 616

Some Personality Correlates of the Relationships Between Individuals and Organizations *H. Russell Johnston, Jr.* 623

Comparison of Job Evaluation Methods: A "Policy-Capturing" Approach Using the Position Analysis Questionnaire *David D. Robinson, Owen W. Wahlstrom, and Robert C. Mecham* 633

Comparisons of Self- and Superior Ratings of Managerial Performance *Herbert G. Heneman III* 638

Effects of Head Restriction on Drivers' Eye Movements and Errors in Simulated Dangerous Situations *John A. Allen, Jr., Stephen R. Schroeder, and Patricia G. Ball* 643

(*Continued on next page*)

(*Continued from previous page*)

SHORT NOTES

Interaction of Subordinate and Leader Characteristics in Moderating the Consideration-Satisfaction Relationship *David M. Herold* 649

Disparagement by a Subordinate as an Influence on the Use of Power *Eugene M. Fodor* 652

Personal Characteristics Differentiating Female Executive from Female Nonexecutive Personnel *Robert F. Morrison and Maria-Luise Sebald* 656

University Climate Perceptions by Black and White Students *C. Michael Pfeifer, Jr., and Benjamin Schneider* 660

LIST OF MANUSCRIPTS ACCEPTED 655

SPECIAL ANNOUNCEMENT 659

INSTRUCTIONS TO AUTHORS

Articles must possess: (*a*) conciseness; (*b*) unambiguous and simple vocabulary; (*c*) conformity to accepted technical style in tables, terminology, and references; and (*d*) conclusions that are clearly related to the evidence presented. Authors must progress from a statement of problem or purpose, through analysis of evidence, to conclusions and implications.

Criteria for evaluation include: (*a*) significance of contribution, (*b*) technical adequacy, (*c*) appropriateness for the *Journal of Applied Psychology*, and (*d*) clarity of presentation.

See the *Publication Manual of the American Psychological Association* (2nd ed.), $3, for the format of manuscripts. Articles not prepared in this manner cannot be reviewed. Note sections on typing (*all copy must be double-spaced*), tables, figures, and references. Do not abbreviate, for example, journal titles in reference lists, the word *subject*, names of groups, etc. Each manuscript copy must include an abstract of 100–120 words typed on a separate sheet of paper.

Optional blind review must be specifically requested when a manuscript is first submitted. The author's name and affiliation must appear only on a separate title page, included with each copy of the manuscript. Footnotes containing the identity of the author or his affiliation must be on separate pages.

Check recent issues of the *Journal of Applied Psychology* for approximate length of regular articles. (One printed page equals three double-spaced manuscript pages.) A few longer articles of special significance are printed as monographs. Short Notes feature brief reports on studies such as those involving some methodological contribution or important replication.

The original drawing of a figure or an 8 × 10 inch glossy print must be submitted with the original copy of the manuscript. Duplicate copies may be photographic or pencil drawn. Figures *must* be hand-lettered.

Supplementary materials formerly deposited with the National Auxiliary Publications Service need no longer be submitted with the manuscript. Retain supporting and raw data for at least 5 years after publication of the article.

NOTICE TO APA AUTHORS

The American Psychological Association announces publication of the second edition of the *Publication Manual* in August 1974. This new edition, which supersedes the 1967 Manual, will be adopted by the 14 APA journals in 1975.

The new Manual is more comprehensive than the previous edition. It updates APA publication policies and procedures and incorporates changes in editorial practice since 1967. For instance, APA now sends many authors their copy-edited manuscripts for review before they are set into type, and some editors now use blind review procedures. The second edition also includes up-to-date statements of the coverage of each APA journal including the *Journal of Experimental Psychology* which will be published in four separate sections in 1975.

The new *Publication Manual* initiates several changes in APA style. These changes are announced in the August 1974 *American Psychologist* and will be introduced in the APA journals in January 1975. During the period of transition to the new style, authors should note that (*a*) all manuscripts published in 1974 will be copy-edited according to the 1967 Manual, (*b*) manuscripts accepted in 1974 and published in 1975 will be copy-edited to conform to the new Manual. Starting in 1975, accepted manuscripts that depart significantly from the Manual will be returned to authors for correction.

Authors will be encouraged by the changes in the second edition. The new APA style simplifies reference forms; eliminates unnecessary underlines, brackets, and other devices; supports appropriate use of "I" and "we"; and generally clarifies typing requirements. Material is arranged for maximum convenience to authors and typists, and all sections are cross-referenced and indexed.

The new *Publication Manual* is available after August 1 for $3. Send orders to APA Publication Sales, 1200 Seventeenth Street, N.W., Washington, D. C. 20036. Orders of $15 or less must include payment unless they are submitted on institutional purchase order forms.

Journal of Applied Psychology
1974, Vol. 59, No. 5, 539-547

EFFECTS OF INTERMITTENT, MODERATE INTENSITY NOISE STRESS ON HUMAN PERFORMANCE[1]

GEORGE C. THEOLOGUS,[2] GEORGE R. WHEATON, AND
EDWIN A. FLEISHMAN[3]

American Institutes for Research, Washington, D.C.

This study evaluated the effects of prolonged exposure to two noise stressors (random and patterned intermittent, 85 decibels) on the performance of three tasks requiring different abilities (Reaction Time, Rate Control, and Time-sharing). The sensitivity of alternate metrics of performance degradation was also evaluated within an analysis of covariance design. Sixty subjects were randomly assigned to the various treatments, and a series of 2 × 2 × 4 (Noise Intensity × Noise Quality × Trial Blocks) covariance analyses were carried out. The effects of random 85-decibel noise on performance depended on the type of task and performance measure. The Reaction Time task was affected, the Rate Control task was not, and the Time-sharing task was affected only after continued exposure to noise. Patterned noise had insignificant effects. Implications for predicting and generalizing stress effects are discussed.

The possible effects of noise in the environment have been a source of increasing concern. Industrial machinery, urban construction, transportation vehicles, and warning devices all contribute to this special kind of environmental pollution. Noise is certainly an irritant which may cause emotional responses, physiological changes, health problems, or loss of sleep (see e.g., Broadbent, 1957; Kryter, 1970).

While the effects of noise on human performance have been studied, generalizations about the effects of noise on specific tasks are hard to come by. Historically, there appear to be two possible reasons for this situation. First, the research may have employed performance tasks with limited diagnostic power, representing inadequately defined types or categories of performance. Second, global and insensitive measurement techniques have been used.

With respect to performance tasks, Alluisi (1967), Fleishman (1967), and Fleishman and Bartlett (1969) have pointed up the need for research using a wide range of standardized behavioral measures which tap a large number of different categories of human performance. Distinctions between mental and motor performance, for example, are not very meaningful and do not reflect our current knowledge about more specific categories of human skills (Fleishman, 1972). Generalizations about the effects of environmental stressors on the various tasks within such all inclusive categories of human performance are likely to be imprecise and difficult to achieve.

Research on a variety of factors affecting human performance repeatedly reveals differential effects for different aspects of performance (Fleishman & Bartlett, 1969). Such studies have been concerned with the effects of stress (Gorham & Orr, 1957), diet (Brozek, Fleishman, Harris, Lassman, & Vidal, 1955), high altitudes (Evans & Consolazio, 1967; Carver & Winsmann, 1968), and drugs (Fleishman, 1967, 1972; Elkin, Fleishman, Van Cott, Horowitz, & Freedle, 1965). For example, in a study (Fleishman, 1967) of the effects of the drug scopolamine, the effect of a specific dosage was a function of the particular ability (e.g., finger dexterity, multilimb coordination, short-term memory) required by the task. It was shown that the absolute performance decrement, the time to

---

[1] This paper is based, in part, on research supported by the National Aeronautics and Space Administration under NASA Grant NAS1-9854 with the American Institutes for Research. Work under this grant was reported in full in Theologus, Wheaton, Mirabella, Brahlek, and Fleishman (1973).

[2] Now with BRX, Inc., Washington, D.C.

[3] Requests for reprints should be sent to Edwin A. Fleishman, American Institutes for Research, 3301 New Mexico Avenue, N.W., Washington, D.C. 20016.

reach this maximum effect, and the time to recover from the effect, depended on the ability measured by the task. It is important to note that all of the aforementioned studies used tasks in which the abilities measured by the task had been empirically determined (Fleishman, 1967).

Appropriate selection of tasks is necessary but not sufficient for thorough assessment of factors affecting performance. Thus, for a meaningful evaluation of noise effects, the performance metric must be considered in addition to the performance tasks employed. A review of the noise stressor research indicates that noise may have its most potent effects on metrics other than mean performance level, percent correct, etc. (Broadbent, 1957). Adequate examination of noise effects, therefore, requires a wider range of metrics capable of detecting other more subtle or complex effects. Chambers (1963), for example, has proposed a stress effects taxonomy which may be adapted to reflect the possible categories of noise effects. For example, his 19 categories may be grouped into five more general classes of performance effects. These five major categories are as follows: (*a*) changes in performance distribution over time, (*b*) decreases in performance rate, (*c*) changes in performance accuracy, (*d*) changes in performance distribution among tasks, among task components and within tasks, and (*e*) increases in performance rate.

The present study is part of a larger effort (Theologus, Wheaton, Mirabella, Brahlek, & Fleishman, 1973) designed to develop a standardized performance battery of tasks for the assessment of the effects of environmental and other stressors on human performance. The rationale of this program is that a combination of (*a*) tasks, representing a range of empirically defined abilities and (*b*) metrics, differentially describing each effect category may provide the basis for a more sensitive and diagnostic methodology with which to establish the behavioral profile of a stressor's effects. Measurement of changes in these components should provide quantitative evaluations of the disruptive effects of the stressor and of man's vulnerability to it. The approach recognizes that while a particular stressor may be studied with respect to one or two modes of performance, it may often modify other aspects as well.

Specifically, in the study reported, the effects of two noise stressors on the performance of three tasks, requiring different abilities, selected from a prototype standardized assessment battery were evaluated. The evaluation was designed to provide information regarding (*a*) the sensitivity of tasks in the battery to two kinds of moderate intensity noise (randomly intermittent and patterned) and (*b*) the sensitivity of alternative metrics of performance degradation. These effects were evaluated over relatively prolonged periods of exposure to noise, compared to the short exposure periods typically studied.

METHOD

Subjects

The 60 subjects employed in this study were male students obtained from area universities. They were paid volunteers, averaging 21 years of age. All subjects were screened for hearing defects.

Apparatus

In earlier work a device was developed (Figure 1) that allowed the assessment of nine separate perceptual–motor human abilities (Parker, Reilly, Dillon, Andrews, & Fleishman, 1965). Briefly, the subject's console contains a cathode ray tube (CRT), a pair of control sticks for one- or two-hand tracking, and a variety of meters and back-lighted push buttons. The experimenter's console contains an array of push buttons and rotary switches for programming experimental tasks. It also includes meters and clocks, and counters for recording subject's responses. For the present research this equipment was placed in two rooms separated by a one-way mirror. The experimenter and his console were located in one room while the subject and his console were in the other.

Abilities Measured

Of the many tasks measurable on this apparatus, three were selected for study. The first experimental task was the *Reaction Time* task. In this task subjects were required to keep the index finger of their nonpreferred hand on a back-lighted push button and to press the button as rapidly as possible whenever the light behind the button came on. The rate of stimulus presentation was eight per minute, with stimulus onsets spaced at approximately equal intervals. This task reflected the Reaction Time ability factor, which is defined as the speed with which an individual can respond to a stimulus when it

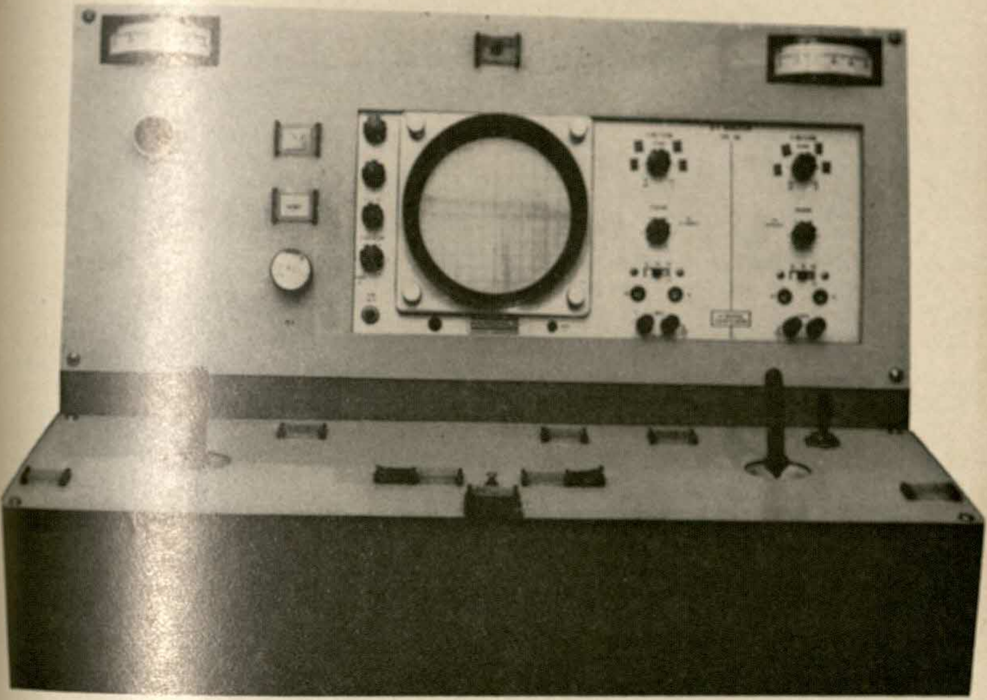


FIGURE 1. The Perceptual-Motor Console (PEMCON), developed to measure separate perceptual-motor abilities. (Originally described in a report by Parker, Reilly, Dillon, Andrews, & Fleishman, 1965. Copyrighted by NASA, 1965.)

appears (Fleishman, 1964). Measures of response latency were recorded on an electronic clock.

The second task was the *Rate Control* task which required subjects to track, in a compensatory fashion, a point of light which moved away from the center of a CRT, under sine-wave, first-order, dynamic programming. To track the dot of light, subjects manipulated, with their preferred hand, a control stick mounted on a universal joint. Movement of the control stick simultaneously affected the x and y axes of the CRT through a pair of independent potentiometers. This task represented the Rate Control factor, which has been defined as the ability to make appropriately timed control adjustments to stimuli varying in rate of movement (Fleishman, 1958). The measure provided was integrated tracking error recorded on galvanometers for the x and y axes.

Lastly, a *Time-sharing* task was employed. This task required subjects to perform the previous two tasks simultaneously. That is, subjects had to perform the Reaction Time task with their nonpreferred hand while simultaneously tracking the point of light with their preferred hand. The use of these two components in the Time-sharing task permitted examination of the effects of noise on the Reaction Time and Rate Control factors both independently and in combination.

Independent Variables

The manipulated variables were noise intensity and noise quality. The noise intensities were 0 decibels (re: .0002 dynes/cm^2) and 85 decibels. Noise quality was represented by randomly intermittent noise and patterned intermittent noise. The random noise consisted of 85 decibels of broadband white noise delivered in bursts of random duration, varying in nine one-second steps between one and nine seconds. The interpolated intervals of silence were likewise random, varying in three one-second steps between one and three seconds. This noise averaged five seconds on and two seconds off. The patterned noise consisted of 85 decibels of broadband white noise presented in five-second bursts with interpolated two-second quiet intervals.

These noise stressors were delivered to subjects through a set of Grason-Stadler earphones (Model TDH-39) from a Viking tape deck (Model 230). Intensity at the earphones was controlled by a Daven step attenuator. The 85-decibel intensity level employed in this study is well within the safe limits at all frequency levels (Kryter, Ward, Miller, & Eldredge, 1966). It has also been shown that the conditions chosen were in fact stressful in that moderate intensity, intermittent white noise has been shown to increase EMG, heart rate, skin conductance, and systolic blood pressure (Glass & Singer, 1972).

TABLE 1

EXPERIMENTAL DESIGN

Task	Subject group	Noise (decibel)	Noise quality	
			Blocks of trials	Blocks of trials
Reaction Time	1	0	Quiet	Quiet
	2	85	Random	Patterned
Rate Control	3	0	Quiet	Quiet
	4	85	Random	Patterned
Time Sharing	5	0	Quiet	Quiet
	6	85	Random	Patterned

Experimental Design

The experimental design is shown in Table 1. In this design the 60 experimental subjects were randomly assigned to one of six groups ($n = 10$). Two groups of subjects performed on each of the experimental tasks. Within each task, one group was assigned to the zero-decibel noise level and the other to the 85-decibel condition. While confounding the patterned random noise manipulation with sequence, this design permitted the most efficient use of the experimental subjects and provided for a reduction of differences among subjects within noise intensity levels.

Procedure

On each day of testing, four subjects reported to the laboratory at half-hour intervals beginning at 8:30 a.m. There were three sessions per day, each lasting approximately one-half hour. There was a 1½-hour break between sessions. Each subject's experimental day lasted approximately 4½ hours.

During the first session each subject was brought individually to his work station, seated at the console, and given instructions relevant to his particular task. During these trials, the subject was instructed to put on the earphones, although no noise stimulus was presented. In this session, 20 trials were administered, each trial lasting one minute, with one-minute rest pauses given at the end of each block of five trials. For the Reaction Time task, eight presentations of the stimulus light were administered within each one-minute trial. In the Rate Control task, tracking performance was continuous over the one-minute trial. From the subject's point of view, performance within and between trials was uninterrupted for five-minute periods.

The second and third testing sessions were used to gather the experimental data. Prior to the second session each subject was assigned either to the experimental condition (85 decibels) or to the control condition (0 decibels). During the second session, each experimental subject was exposed to the random noise while performing his particular task. During the third session, the patterned noise was presented. Control subjects received no noise stimulation in either the second or third sessions, although they were required to wear the earphones. The testing procedures for the second and third sessions were identical to those for the first session except that two one-minute practice trials were administered prior to the 20 trials, and these 20 trials were massed.

Dependent Variables

This experiment produced the following basic data. For each subject assigned to the Reaction Time and Time-sharing tasks, 160 response latencies measured in milliseconds were recorded for each experimental session. For each subject who performed on the Rate Control and Time-sharing tasks, 20 integrated error scores were recorded in each experimental session for the x and y tracking axes. These basic data were not subjected to analysis. Rather, they provided a basis for derivation of dependent measures specifically tailored to provide information on the differential sensitivity of selected performance measurement categories.

To provide more stable estimates of subjects' performance, the data within each session were collapsed into four blocks of five trials. All subsequent measures were derived from these blocked data.

One of the primary reasons for the research was to determine whether the performance measurement categories would exhibit differential sensitivity to noise. Consideration of the basic data suggested a variety of dependent measures which could be used to represent each of the five major categories which we imposed on Chamber's (1963) degradation taxonomy. It was decided to limit initial concern to those rather commonly used to describe performance on reaction time and tracking tasks such as those used in this study. The categories selected and the measures representing them were as follows:

Category 1. Changes in performance distribution over time; assessed by (*a*) frequency of response blocks and (*b*) mean duration of response blocks.

Category 2. Decreases in performance rate; assessed by (*a*) mean response latency and (*b*) reaction time.

Category 3. Changes in performance accuracy; assessed by (*a*) total mean integrated tracking error, (*b*) x axis mean integrated error, (*c*) y axis mean integrated error.

The derivation of these seven measures from the obtained data was as follows.

The first dependent measure used was *mean response latency* for each block of trials. This measure is conceptually independent of the usual reaction time measure. While it includes data which would normally be thought of as reaction time, it also includes longer latencies reflecting time to detect the stimulus. Thus, some subjects may be capable of responding quickly, but may fail to perceive the stimulus when it appears. (This distinction is especially critical in the Time-sharing task.) To normalize these distributions reciprocal transformation of the data was employed.

The appearance of longer than normal response latencies in the data suggested that, in addition to variation in response latency, a lapse-in-performance phenomenon was being measured. It was assumed that these longer responses were related to momentary lapses in attention during which the subject failed to detect presentation of the stimulus. This phenomenon is referred to in the literature as response blocking (Bills, 1937). Since no universally acceptable quantitative definition of a response block is presently available, an operational definition which had proved useful in a study of long-term visual monitoring (Baker & Theologus, 1972) was adopted. Therefore, for the present study, a response block was defined for each subject as any latency which was two or more times longer than the mean of h:s 20 fastest response latencies during the familiarization session. Given a definition of a response block for each subject, two measures were derived: (*a*) The number of response blocks within each block of trials was counted to develop the measure of *frequency of response blocks*. (*b*) The *mean duration of response blocks* was calculated within each block of trials.

Examination of these three derived measures revealed that the measure of *mean response latency* evaluated the total distribution of scores for each subject while the measures of *frequency of response blocks* and *mean duration of response blocks* evaluated scores in the upper tail of the distribution. A fourth measure (representing the subject's reaction time) was developed. Since all scores two or more times longer than the mean of the subject's 20 fastest response latencies were defined as response blocks, we defined all of the scores falling below this criterion as reaction times. Therefore, for each subject the mean of all scores less than a response block was calculated within each block of trials to generate the measure of *mean reaction time*.

The error data obtained on the Rate Control and Time-sharing tasks were used to obtain three measures within each block of trials including *total mean integrated error* on both tracking axes, *mean integrated error on the* x *axis*, and *mean integrated error on the* y *axis*.

Data Analysis

The primary analytic technique was analysis of covariance. For each of the performance measures, the covariate employed to adjust the measures from the two experimental sessions was the similar measure calculated for each subject on the last 10 trials from the familiarization session. Fourteen 2 × 2 × 4 (Noise Intensity × Noise Quality × Blocks) designs were analyzed, followed by single degree of freedom *F* tests calculated on the adjusted means derived from the analyses of covariance.

RESULTS

The more salient and interesting results are presented in Figures 2 through 8. In these figures performance curves are shown for different performance metrics and the three kinds of performance tasks which were studied. Performance is shown as a function of noise intensity (experimental = 85 decibels, control = 0 decibels), session (Session 2 = random intermittent noise, Session 3 = patterned intermittent noise), and trial block within session. Effects due to the first two of these variables were of primary interest.

Intensity Effects

While in most of these figures the performance of the random 85-decibel noise group is shown to be inferior to that of the control group, the overall mean effects as tested in the complete analyses of covariance did not reach statistical significance. However, a priori comparisons between the experimental (85 decibels) and control (0 decibels) groups at each block of trials, especially in the random intermittent noise sessions, did reveal a number of significant effects.

As shown in Figure 2, experimental subjects who were exposed to 85-decibel random intermittent noise while performing on the Reaction Time task had longer latencies than did control subjects at the first ($F = 7.12$, $df = 1/143$, $p < .01$) and second ($F = 6.90$, $df = 1/143$, $p < .01$) blocks of trials. Similar effects on performance of this task were

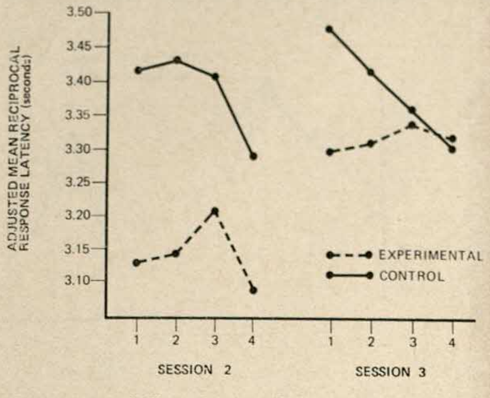


FIGURE 2. Adjusted transformed mean response latency at each block of trials in both sessions on the Reaction Time task.

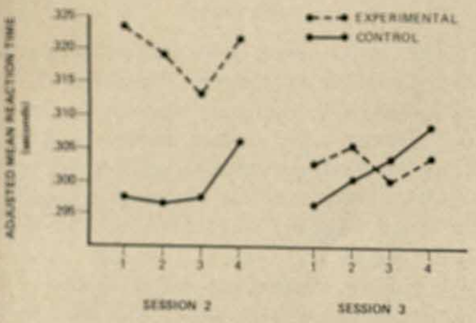


FIGURE 3. Adjusted mean reaction time at each block of trials in both sessions on the Reaction Time task.

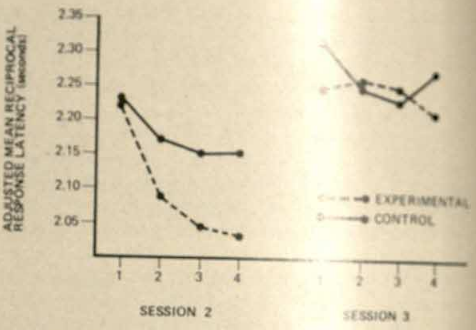


FIGURE 5. Adjusted transformed mean response latency at each block of trials in both sessions on the Time-sharing task.

found for the reaction time measure; the curves in Figure 3 indicate that mean reaction time of the experimental groups, performing under 85-decibel random intermittent noise, was slower than that of the control group (0 decibels) at both the first ($F = 8.06$, $df = 1/143$, $p < .01$) and second ($F = 6.36$, $df = 1/143$, $p < .05$) blocks of trials.

In contrast to the latency results with the Reaction Time task the compensatory tracking data from the Rate Control task showed no difference as a function of noise intensity. This outcome is clearly shown in Figure 4 for total mean integrated tracking error. The x-axis and y-axis data, when considered independently, parallel these functions and support the absence of an effect.

Finally, the data presented in Figures 5 and 6 for the Time-sharing task parallel, to a degree, those presented for the Reaction Time task. While the results do not reach conventional levels of significance, they suggest that initial exposure to noise increases both response latency and reaction time scores on the Time-sharing task, relative to performance under the no-noise control condition.

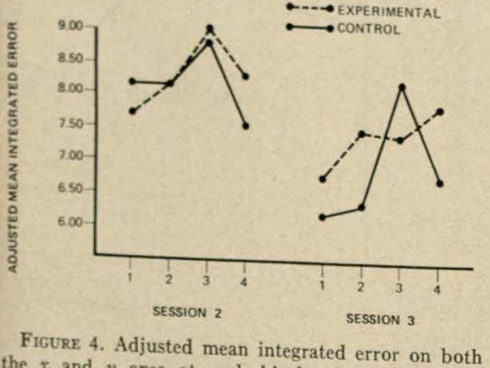


FIGURE 4. Adjusted mean integrated error on both the x and y axes at each block of trials in both sessions on the Rate Control task.

Session Effects

It should be recalled that there were three sessions; the first familiarized subjects with the task; the second presented the random intermittent noise; and (after a 90-minute break) the third session presented the patterned intermittent noise. Figures 2 through 8 include the second and third sessions, during which the experimental group was exposed to the noise conditions. The effects reported in this section thus may be viewed as due either to the quality of noise (random intermittent versus patterned intermittent noise) or to the first versus second exposure to noise.

Significant interactions between noise intensity and session were obtained on the Reaction Time task. As shown in Figures 2 and 3, both response latency ($F = 9.06$, $df = 1/18$, $p < .007$) and reaction time ($F = 9.89$, $df = 1/18$, $p < .006$) decreased across sessions, with the experimental group clearly showing the greater improvement.

For the Rate Control task neither the interaction between noise intensity and session nor the main effects of session reached statistical significance (Figure 4).

Significant main effects due to sessions were obtained on the Time-sharing task for the measures of response latency (Figure 5), frequency of response blocks (Figure 7), and total integrated error (Figure 8). Mean response latency ($F = 6.64$, $df = 1/18$, $p < .02$) decreased across sessions as did the frequency of response blocks ($F = 10.30$, $df = 1/18$, $p < .005$), while total integrated tracking error ($F = 5.42$, $df = 1/18$, $p < .03$) improved. However, while greater improvement across the two sessions seemed to be associated with the experimental group (85 decibels), session by noise intensity interactions were not significant.

Taken together these results with regard to session effects on the Time-sharing task indicate some improvement in the second session regardless of whether noise was present.

The data in Figures 2 through 8 also suggest that unlike the results for the session involving random intermittent noise, where effects were obtained, the group under patterned noise never performed worse than the quiet control group. The performance curves for the patterned noise and quiet groups overlap regardless of task or measure. This appears to be independent of how disruptive the random noise had been during the preceding session.

Discussion and Conclusions

The results provide support for the approach taken to assess stress effects. This approach is predicated on the rationale that different types and levels of stressors will variously affect different components of human task performance and that these effects may become manifest only in certain categories of performance degradation.

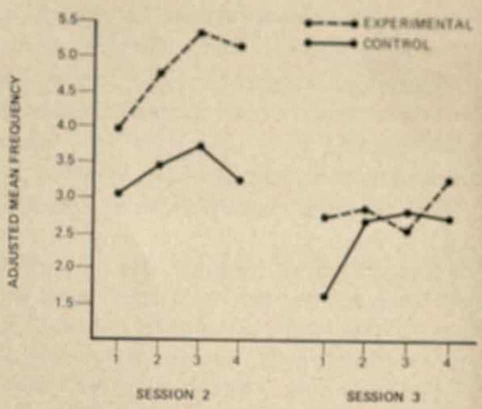


Figure 7. Adjusted mean frequency of response blocks at each block of trials in both sessions on the Time-sharing task.

In the present experiment the use of tasks known to sample different abilities (selected from a standardized assessment battery), coupled with measures from a stress-effects taxonomy, permitted demonstration of differential noise effects. For example, performance on the Reaction Time task was initially degraded following exposure to randomly intermittent noise. By using a degradation taxonomy, however, it was possible to show that the degradation was highly specific. As evidenced by the response latency and reaction time data, initial exposure to noise slowed the speed at which subjects normally re-

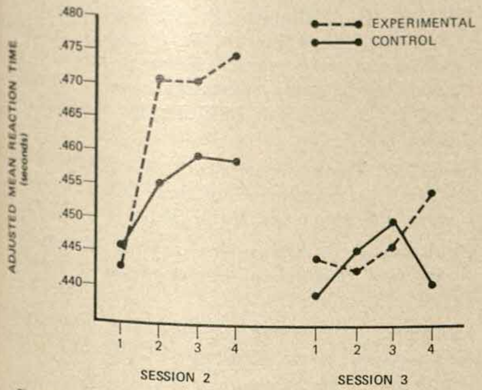


Figure 6. Adjusted mean reaction time at each block of trials in both sessions on the Time-sharing task.

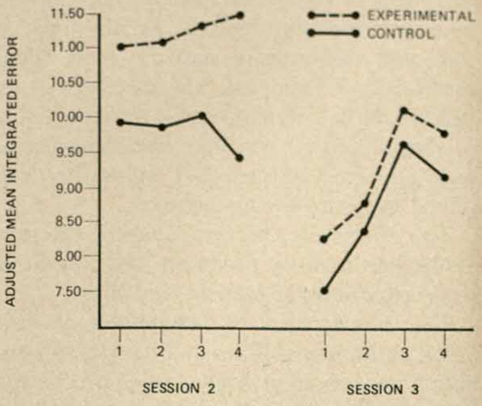


Figure 8. Adjusted mean integrated error on both the x and y axes at each block of trials in both sessions on the Time-sharing task.

sponded, but did not increase the number or duration of response blocks. In other words, a Category 2 degradation (i.e., decreases in performance rate) occurred, while a Category 1 decrement (i.e., a change in performance distribution over time) was not observed.

The potential value of using different tasks to assess stress effects is illustrated by the total integrated tracking error data which were analyzed for both the Rate Control and Time-sharing tasks. As already noted (Figure 4) there clearly was no degradation in Category 3 (i.e., changes in performance accuracy) on the Rate Control task. This does not appear to be the case, however, when comparable data from the Time-sharing task are examined. Although the differences shown in Figure 8 do not reach conventional levels of significance, they do suggest that Category 3 degradation is taking place, especially under the random intermittent noise conditions of Session 2.

These same data, when considered jointly with Category 1 (Figure 7) and Category 2 (Figures 5 and 6) data, strongly suggest a different time course for the emergence of effects on Time-sharing performance. The overall impression is that degradation is becoming more pronounced with the passage of time. Had the duration of the initial exposure been increased, degradations in all three categories might have occurred during Time-sharing performance.

Taken together, the results indicate that the effect of 85-decibel *random* noise on human performance depends on the type of task and performance measure used. Thus, the Reaction Time task was affected, but the task requiring Rate Control ability (timing of control action) was not. The Time-sharing task appeared affected but only after continued exposure to this noise.

The effects of *patterned* noise of the same 85-decibel intensity produced insignificant effects, regardless of task or measure.

The combination of taxonomies of tasks based on human abilities and metrics of stress effects appears to provide a more precise way of predicting noise effects. The specificity of the current results with regard to both components of performance and degradation categories underscores the complexity of noise effects. Although further research is required to confirm these results, the obtained differential sensitivity of measures and tasks indicates the potential of the approach. Study of a wider range of stressors, measures, and tasks may permit development of a standard method for diagnosing stress effects and for deriving behavioral profiles for a wide variety of stressors.

REFERENCES

Alluisi, E. A. Methodology in the use of synthetic tasks to assess complex performance. *Human Factors*, 1967, **9**, 375–384.

Baker, W. J., & Theologus, G. C. Effects of caffeine on visual monitoring. *Journal of Applied Psychology*, 1972, **56**, 422–427.

Bills, A. G. Blocking in mental fatigue and anoxemia compared. *Journal of Experimental Psychology*, 1937, **20**, 437–452.

Broadbent, D. E. Effects of noise on behavior. In C. M. Harris (Ed.), *Handbook of noise control*. New York: McGraw-Hill, 1957.

Brozek, J., Fleishman, E. A., Harris, S., Lassman, F. M., & Vidal, J. H. Sensory functions and motor performance during maintenance of survival rations. *American Psychologist*, 1955, **10**, 502.

Carver, R. P., & Winsmann, F. R. Effect of high elevation upon physical proficiency, cognitive functioning, and subjective symptomatology. *Perceptual and Motor Skills*, 1968, **26**, 223–230.

Chambers, R. M. Operator performance in acceleration environments. In N. M. Burns, R. M. Chambers, & E. Hendler (Eds.), *Unusual environments and human behavior*. London: Glencoe, 1963.

Elkin, E. H., Fleishman, E. A., Van Cott, H. P., Horowitz, H., & Freedle, R. O. *Effects of drugs on human performance: Research concepts, test development, and preliminary studies*. Washington, D.C.: American Institutes for Research, October 1965.

Evans, W. O., & Consolazio, F. C. Effects of high altitude on performance of three different types of work. *Perceptual and Motor Skills*, 1967, **25**, 41–50.

Fleishman, E. A. Dimensional analysis of movement reactions. *Journal of Experimental Psychology*, 1958, **55**, 430–453.

Fleishman, E. A. *The structure and measurement of physical fitness*. Englewood Cliffs, N.J.: Prentice-Hall, 1964.

Fleishman, E. A. Performance assessment based on an empirically derived task taxonomy. *Human Factors*, 1967, **9**, 349–366.

Fleishman, E. A. On the relation between abilities, learning, and human performance. *American Psychologist*, 1972, 27, 1017–1032.

Fleishman, E. A., & Bartlett, C. J. Human abilities. *Annual Review of Psychology*, 1969, **20**, 349–380.

Glass, D. C., & Singer, J. E. Behavioral aftereffects of unpredictable and uncontrollable aversive events. *American Scientist*, 1972, **60**, 457–465.

Gorham, W. R., & Orr, D. B. *Research on behavior impairment due to stress.* Washington, D.C.: American Institutes for Research, September 1957.

Kryter, K. D. *The effects of noise on man.* New York: Academic Press, 1970.

Kryter, K. D., Ward, W. D., Miller, J. D., & Eldridge, D. H. Hazardous exposure to intermittent and steady-state noise. *Journal of the Acoustical Society of America*, 1966, **39**, 451–464.

Parker, J. F., Reilly, R. E., Dillon, R. F., Andrews, T. G., & Fleishman, E. A. *Development of tests for measurement of primary perceptual-motor performance.* (NASA Contractor Report 335) 1965. (NTIS No. N66-12127)

Theologus, G. C., Wheaton, G. R., Mirabella, A., Brahlek, R. E., & Fleishman, E. A. *Development of a standardized battery of performance tests for the assessment of noise stress effects.* (NASA Contractor Report CR-2149) 1973. (NTIS No. N73-17102)

(Received August 5, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 5, 548–554

EFFECT OF NOISE ON INTELLECTUAL PERFORMANCE[1]

NEIL D. WEINSTEIN[2]

University of California, Berkeley

The effect of realistic noise levels on an intellectually challenging task was examined. Thirty-three college students performed proofreading in quiet or with a background of intermittent teletype noise (70 decibels). As predicted, noise subjects did not differ significantly from quiet subjects in detecting spelling errors, but were poorer at identifying grammatical errors ($p <$ .0[illegible]). Contrary to expectations, recall of the content of the proofreading passage was unaffected. Detailed analysis revealed that subjects initially worked more slowly ($p < .01$) and less steadily ($p < .002$) during noise bursts than during intervening quiet periods, but more accurately ($p < .01$). The results are interpreted in terms of the feedback available to the subjects and the strategies they adopt to cope with the distraction.

Early studies of the effects of noise on complex activities like learning, reading comprehension, and intelligence test performance were often frustrated by inconclusive results (reviewed in Broadbent, 1957). Consequently, much recent research on noise has emphasized simpler tasks chosen for theoretical clarity and intended to tap more fundamental cognitive processes. Vigilance, response time, visual search speed, and the ability to solve simple arithmetic problems have been frequently tested (Broadbent, 1971; Kryter, 1970). Still, the outcome of these studies, taken as a whole, has been no less confusing than the results of earlier research, and critics (Kryter, 1970) continue to maintain that noise poses no threat to any kind of nonauditory performance.

There are at least two possible explanations for the disappointing findings of recent experiments. One explanation is that noise affects only tasks which are intellectually demanding. Certainly, tasks which measure response time and visual search speed place few burdens on intellectual functioning (see Fleishman, 1967, for a discussion of task dimensions).

Another possibility is that the tasks in these noise experiments are so straightforward and unidimensional that the best strategies to use are self-evident. These strategies may leave all *measured* aspects of the subject's performance unaffected, but in some way he must pay in order to maintain speed and accuracy.

Support for this second argument comes from a group of studies which found that noise caused performance to deteriorate if subjects were required to carry out several tasks simultaneously (Finkelman & Glass, 1970; Hockey, 1970a, 1970b; Woodhead, 1966). Such studies have demonstrated that subjects sometimes maintain performance on one activity at the cost of poorer performance on others, but the use of rather exotic tasks and high noise levels (usually above 90 decibels) suggests that the effects of noise on more familiar activities are insignificant, at least in a laboratory setting.

The experiment described in this study was undertaken with two goals in mind: (*a*) to determine whether noise levels far lower than those generally employed would interfere with a familiar task which has some real-life significance; and (*b*) to understand how subjects cope with noise by a more thorough monitoring of performance than is usually attempted.

Proofreading was chosen as the experimental task because it is intellectually demanding, relatively familiar, and because it is sufficiently complex so that the optimum strategy is not obvious. It was predicted that errors such as misspellings, which can be detected by examining a single word, would be little affected by noise. These were viewed as problems of pattern recognition, making relatively few demands on the subject's information processing capabilities. In contrast to these *noncontextual errors*, the investigator expected that *contextual errors*, a category consisting of mistakes like grammatical errors that can be detected only by placing a word in context, would be significantly affected. Subjects would presumably find it difficult to integrate individual words into a coherent structure. It was anticipated that comprehension, because of its incidental character and its presumed difficulty would be impaired most by noise.

---

[1] The author is indebted to Richard S. Lazarus for his advice during the planning of this study and to the National Institute of Mental Health for providing support with a postdoctoral fellowship.

[2] Requests for reprints should be addressed to Neil D. Weinstein, who is now at the Department of Human Ecology and Social Science, Cook College, Rutgers University, P.O. Box 231, New Brunswick, New Jersey 08903.

On the basis of research by many others, noise was not expected to change overall proofreading speed. Still, the mean speed might be concealing a number of interesting effects, so the experiment was designed to make it possible to measure the length of time spent on each line.

METHOD

Subjects

The subjects in this experiment were college students from an introductory psychology class who participated to fulfill a course requirement. Random assignment resulted in 18 subjects in the noise condition and 15 subjects in the quiet condition. Subjects in each group were nearly equally divided between males and females.

Proofreading and Comprehension Tasks

Three proofreading texts were prepared from nonfiction sources. Selections were long enough so that no subject could complete them in the time allotted. Errors were introduced at irregular intervals; on the average, one line in three contained errors. As many as three errors could appear in a single line, but most incorrect lines had only one error. Except for minor variations, the pattern of errors was repeated every 50 lines. All such errors were chosen to fit into one of two mutually exclusive categories, and errors from the two different categories were never present in the same line. *Contextual errors* included errors in grammar, missing words, and incorrect or inappropriate words (e.g., the word *an* appearing instead of the word *and*). Such errors could not be detected by examining each word separately, but only by placing the words in context. *Noncontextual errors* included misspellings and typographical errors. In each 50 lines of text there were seven lines containing contextual errors and eleven lines with noncontextual errors.

Each proofreading passage was typed on a continuous roll of paper in a single column approximately 60 characters (5 inches) wide. The paper was presented to the subject on a typewriter. A mask on the typewriter had a narrow slit which allowed the subject to see only one line at a time. The subject was instructed to press a button in his right hand for each error he found. Finishing a line, he would push the typewriter's carriage-return lever to advance the paper to the next line. Since the subject was unable to look ahead or look back, reading in this manner was probably more difficult than normal, but all subjects appeared to adjust to the procedure very quickly.

Two comprehension tests, one emphasizing recognition of specific topics and phrases, the other concerned with an understanding of the concept and arguments presented in the selections, were prepared for each of the last two passages.

Noise

The noises heard by the subject were the normal, unamplified sounds of a teletype machine (Teletype Corporation Model 33ASR) placed alongside the cubicle in which the subject worked. The noise was produced in irregularly scheduled bursts lasting from 2½ to 15 seconds. These were arranged so that noise was present for a total of 30 seconds in each one-minute period. Both typing at the teletype keyboard and operation of the teletype's paper tape reader were used so that the background sounds would resemble normal teletype noise. At the subject's ears, the A-weighted sound level was 66 decibels for typing and 70 decibels for the paper tape reader. In between noise bursts, one could hear the hum of the teletype motor, 55 decibels. Subjects in the quiet condition worked with ambient background noise of approximately 36 decibels. (The C-weighted values of these same sounds were 71, 75, 68, and 55 decibels, respectively.)

Design and Procedure

The experiment was presented to the subject as a study of the relationship between selected aspects of intellectual performance and a number of personality traits that were to be tested in a later session. Written instructions described the proofreading task and gave examples of all types of proofreading errors. No distinction was made between contextual and noncontextual errors. Subjects were not informed that their comprehension would be tested later. Although each subject could work at his own rate, the experimenter stated that both speed and accuracy were important and were given equal weight. The subject was then seated at the typewriter in the cubicle and was permitted to practice proofreading on several sample lines.

The time the subject spent on a line and the number of errors he detected in it were recorded on a multichannel recorder. The output of a micro-

TABLE 1

PROOFREADING PERFORMANCE OF NOISE AND QUIET GROUPS ON PASSAGES TWO AND THREE

Group	Dependent variables					
	2[b]	3	2	3	2	3
	Mistakes on noncontextual error lines[a] (% wrong)		Mistakes on contextual error lines[a] (% wrong)		Response time (second/line)	
	Means[c]					
Quiet	28	28	40	44	4.56	4.27
Noise	36	35	58	65	4.68	4.09
	Variances[c]					
Quiet	90	120	100	90	.18	.38
Noise	160	300	240	430	1.16	.56

[a] Proportion of lines on which mistakes were made.
[b] Passage number.
[c] Based upon scores subjected to a linear covariance adjustment using the performance on passage one as the covariate.

phone placed near the noise source was recorded simultaneously on a third channel so that lines worked in noise could be differentiated from lines worked in quiet.

All subjects worked six minutes in quiet on the first reading selection. Each subject's proofreading accuracy and his speed during the last three minutes of the selection were employed as covariates when an analysis of covariance was used to compare the performances of the two groups during the remainder of the session.

After being told to stop, each subject rested several minutes while the experimenter put the next proofreading selection in place. Before beginning again, subjects in the noise condition were told: "You will hear some teletype noise in the background. Just ignore it and keep on working." Seven minutes were allowed for work on each of the second and third passages, with a short rest in between. Thus, the quiet group worked in quiet on all three passages; the noise group worked in quiet on the first passage and with noise on the last two. The selections were always presented in the same order.

At the conclusion of the proofreading, the subject was told that the experimenter would like to find out how much the subject could recall of the content of the passages. The subject then worked on the four comprehension tests in a fixed order, attempting only those questions dealing with information presented in the lines he had finished reading. The experimental session ended with a brief, structured interview which dealt with the subject's evaluation of his performance and with his reactions to the noise.

RESULTS

Comparison of the Noise and Quiet Groups

The proofreading speed and accuracy of the noise and quiet groups were compared by analysis of covariance, and a *t* test was used to examine comprehension in the two conditions.[3] Accuracy was measured in terms of the proportion of contextual error lines or noncontextual error lines which were incorrect. A line was considered incorrect if the number of errors reported by the subject differed by one or more from the actual number. The great majority of mistakes were errors of omission. Table 1 presents group means and variances for accuracy and speed on each of the last two passages. There were no significant differences between the two groups on the first, control passage for any dependent variable.

As anticipated, the effect of noise on noncontextual errors was not statistically significant. The noise group did overlook approximately 25% more errors on both passages, but the variation among individual performances within each group was so large that the power of the tests was only about .25. The effect of noise on contextual errors also agreed with the initial prediction. The noise group was almost 50% less accurate, a difference which is significant at the .001 level for both passages, and no improvement was apparent from the second passage to the third. Roughly 30% of the total variance in contextual error accuracy could be attributed to the experimental manipulation (see Fleiss, 1969, for the formula used for this measure). A *t* test confirmed that the effect of noise on contextual errors was significantly greater than the effect on noncontextual errors ($p <$ .05 and $<$ .02 on the second and third passages, respectively).

---

[3] All tests reported in this study are two-tailed. Statements referring to the power of the statistical tests reveal the probability that an F or t value would prove significant beyond the .05 level if noise actually produced a relative change in performance of 20%. The figure of 20% was chosen to illustrate the ability of the experiment to detect a fairly substantial change.

There were no changes produced by noise on any of the four comprehension tests. The quiet group correctly answered an average of 60.6% of the questions from the combined tests on the second passage and 65.8% of the questions from the third passage, whereas the corresponding values for the noise group were 61.4% and 67.6%. The internal reliability of the comprehension tests averaged .80, while the power of the statistical comparisons was about .75 for both selections.

The last column in Table 1 shows that there were no significant effects on mean proofreading speed, in spite of the fact that the power of the tests was relatively high (over .8 and .9 on the last two passages).

Additional data presented in Table 1 reveal that the noise group was more variable in all instances.[4] The difference between groups was statistically significant for contextual error accuracy on the third passage ($p < .01$) and for response time on the second passage ($p < .002$), indicating that individual differences in response to noise may exist even in the absence of any main effect.

Noise Subjects During Periods of Noise and Quiet

A more detailed analysis of the noise group compared proofreading performance during bursts of noise with proofreading during periods of relative quiet.[5] Significance levels were computed using t tests for correlated means ($df = 17$). As presented in Table 2, significantly fewer mistakes were committed on noise lines than on quiet lines during the second passage ($p < .01$), but this effect had diminished by the third passage and was no longer statistically significant.

TABLE 2

PERFORMANCE OF NOISE SUBJECTS ON NOISE AND QUIET LINES OF PASSAGES TWO AND THREE

Lines	Dependent variables					
	2[b]	3	2	3	2	3
	Mistakes on error lines[a] (% wrong)		Response time (second/line)		Standard deviation of response time[c] (second/line)	
Quiet	56	52	3.87	3.61	1.15	1.17
Noise	46	48	4.37	3.82	1.57	1.38

[a] Mean proportion of lines containing either contextual or noncontextual errors on which mistakes were made.
[b] Passage number.
[c] Group mean of the standard deviation of response time calculated for each subject.

By recording the length of time spent on each line, it was possible to determine whether subjects were faster when noise was on or off and whether their response time was as stable in noise as in quiet. The values displayed in Table 2 show that during the second passage the mean response time and the standard deviation of the response time were considerably greater on lines worked with noise than on lines worked in quiet (significant at the .01 and .002 levels, respectively). Similar, but smaller, effects appeared on the third passage, with both means and standard deviations approaching significance at the .1 level. In spite of these differences, when all lines are pooled, a noise subject exhibits slightly less variability than does an average quiet subject (significant only on the third passage, $p < .05$). The trend toward both equal speed and equal variability during noise and quiet, apparent in Table 2, was also significant ($p < .05$ and $p < .10$ for means and standard deviations, respectively).

Subject Interviews

From interviews at the end of each session, it became clear that subjects did not believe that the noise had interfered significantly with their proofreading. While practically all agreed that the noise was distracting at first, they felt that they had adapted to it. Aside from the frequent comment, "I tried to concentrate more," subjects were unaware of any alteration in their proofreading strategy. In general, changes in speed were mentioned more often than changes in accuracy.

---

[4] Although these differences violate the assumption of homoscedasticity underlying the analysis of covariance, the results already discussed are sufficiently clear-cut that a correction for unequal variances would not change any of the conclusions.

[5] A *noise line* was defined as one in which noise from typing or from the paper tape reader was present at least half of the time. A *quiet line* was defined as one in which no noise was present. The exclusion from the former category of lines with only a small portion of noise was necessary to avoid a subtle artifact which would otherwise appear if the speeds on noise and quiet lines were compared (Fisher, 1972).

DISCUSSION

Although the original hypotheses of this study concerning the effects of noise on proofreading and speed were confirmed, the expectation that comprehension would show the greatest deterioration was proven incorrect. Apparently, subjects can combine individual words into a meaningful structure and still be unable to detect contextual errors. It seems that several different processes must be called upon to account for the results.

The interviews revealed that subjects were very poor sources of information concerning their proofreading accuracy. In the absence of any other feedback they may be forced to use their degree of understanding of the passage as the criterion by which to judge whether they are reading carefully enough to locate mistakes. Subjects may respond to noise by attempting to maintain their level of comprehension, without realizing that they may be failing to test the sentences for proper syntax, or may be reading what they expect to find written, unintentionally correcting the errors.

There are several possible reasons why contextual errors were affected more by noise than were noncontextual errors. One explanation stresses the differences in the cognitive processes required to detect such errors. Presumably the pattern recognition skills involved in identifying noncontextual errors are less sensitive to distraction than are the more complex processes involved in contextual error recognition. A competing explanation emphasizes subjectively perceived differences in the frequency of the two types of errors. Because they appear to be considerably more frequent, subjects may concentrate their effort on noncontextual errors. Though both these mechanisms are consistent with the present findings, a third possibility, that difficulty alone is the crucial variable, must be considered. Examination of individual lines, however, showed that for both types of errors the effect of noise was no greater on difficult lines than on easy lines.

In contrast to the inaccuracy of subjects' reports concerning their proofreading success, their comments about their rate of work were fairly accurate. Both the initial disruption and the eventual adaptation they mentioned reflect faithfully the effects on speed and variability presented in Table 2. The ability of individuals to judge how fast they are actually working and their desire to resist any changes in this rate of work may explain why effects of stress on overall speed are seldom observed.

Several other experimenters (Fisher, 1972; Woodhead, 1964) have noticed within-subject differences in performance during intermittent periods of noise and quiet without any overall differences between noise and quiet groups. These observations probably result from the greater statistical power of the within-subject analyses. In the present investigation, for example, the correlation between the speed on noise and quiet lines in the same passage was .98, making the *t* test for correlated means very powerful indeed.

Are any of the theories about noise or stress present in the literature helpful in explaining why one response measure rather than another was affected in this experiment? One theoretical position holds that noise produces an increase in arousal and that it is overarousal which interferes with performance (Broadbent, 1971). The appeal to a change in an undifferentiated bodily state, however, seems quite incapable of dealing with the relatively complicated pattern of heterogeneous effects observed in this experiment.

More useful is Easterbrook's (1959) theory of attentional narrowing, which differentiates between focal and incidental tasks (see Kahneman, 1973). Easterbrook suggests that stress narrows the range of cues which an organism can use, leading to poorer performance on incidental activities. In the present context *focal* means primary, and *incidental* means secondary, or less important, in terms of the subject's priorities. Easterbrook and Kahneman believe that attentional narrowing is caused by physiological arousal due to emotion, stress, or drugs. It is arguable, however, whether mild but distracting noises produce their effects through a large change in the level of arousal. Another possibility is that the change in the deployment of attention is primarily a voluntary process. The subject attempts to compensate for his involuntary attention to the distracting stimulus in order to keep his performance from deteriorating. The new distribution of attention and effort is determined by his beliefs concerning the best strategy to use in overcoming the interference. The greater variability of the noise group in the present investigation, for example, is probably due in part to individual differences in the method chosen to deal with the distraction.

In experiments with multiple tasks, the one most impaired by stress frequently appears to be the one considered least important by the subject. Such subjectively perceived differences in importance can be established explicitly via instructions (Hockey, 1970a; Woodhead, 1966) or implicitly via length of practice, greater novelty of a particular task, need for more frequent responses (Bahrick, Rankin, & Fitts, 1952), differences in feedback, or the subject's inherent cognitive style (Broverman & Lazarus, 1958).

The preceding discussion suggests that the effect of noise on the performance of complex tasks cannot be understood without assigning an active role to the subject. The subject decides which strategy he will use on the basis of whatever information is available to him (e.g., Hockey, 1970b). Theories which treat the subject as a passive figure or hold that the effects of stress result merely from a change in the physiological state of the body ignore the large differences in performance which can be produced by a small change of instructions or feedback.

The interviews and the detailed study of the noise group provide additional support for this point of view. It may be possible to explain the slower and more irregular speed during noise bursts in terms of a passive process such as a brief failure of attention. But it seems difficult to explain the superior accuracy on noise lines without assuming that subjects are aware of the distraction and resist the threatened change in their performance by taking greater care during noise than during quiet.

Why has this investigation produced clear-cut evidence of noise interference when others have frequently yielded small or equivocal results? It appears that the effects observed in the present experiment are due to the task chosen and to the simultaneous monitoring of several different aspects of the subjects' performance, rather than to some peculiar feature of the noise source or subject sample. Because the task is complex and no feedback on accuracy is available, subjects may have made a strategic mistake in responding to the noise. Such a mistake could be corrected readily with appropriate feedback. It is also possible, however, that noise interferes more directly with the ability to proofread and that it would be difficult for subjects to improve their performance. The present study cannot establish which of these two alternatives is most correct.

In spite of current interest in noise as one aspect of a deteriorating environment, the use of realistic noise levels and meaningful tasks is still uncommon. Without theories capable of making accurate generalizations from one situation to another, the use of laboratory analogues must be recognized as a hazardous practice. Investigators interested in the practical applications of their research must attempt to design experiments which duplicate, as nearly as possible, conditions as they occur in real life.

REFERENCES

Bahrick, H., Rankin, R. E., & Fitts, P. M. *The effect of motivation upon peripheral perception during the performance of a central psychomotor task.* (Research Bulletin 52-14) Lackland Air Force Base, San Antonio, Tex.: Human Resources Research Center, 1952.

Broadbent, D. E. Effects of noise on behavior. In C. M. Harris (Ed.), *Handbook of noise control.* New York: McGraw-Hill, 1957.

Broadbent, D. E. *Decision and stress.* London: Academic Press, 1971.

Broverman, D. M., & Lazarus, R. S. Individual differences in task performance under conditions of cognitive interference. *Journal of Personality,* 1958, **26**, 94–105.

Easterbrook, J. A. The effect of emotion on cue utilization and the organization of behavior. *Psychological Review,* 1959, **66**, 183–201.

Finkelman, J. M., & Glass, D. C. Reappraisal of the relationship between noise and human performance by means of a subsidiary task measure. *Journal of Applied Psychology*, 1970, **54**, 211–213.

Fisher, S. A "distraction effect" of noise bursts. *Perception*, 1972, **1**, 223–236.

Fleishman, E. A. Development of a behavior taxonomy for describing human tasks: A correlational–experimental approach. *Journal of Applied Psychology*, 1967, **51**, 1–10.

Fleiss, J. L. Estimating the magnitude of experimental effects. *Psychological Bulletin*, 1969, **72**, 273–276.

Hockey, G. J. R. Effect of loud noise on attentional selectivity. *Quarterly Journal of Experimental Psychology*, 1970, **22**, 28–36. (a)

Hockey, G. J. R. Signal probability and spatial location as possible bases for increased selectivity in noise. *Quarterly Journal of Experimental Psychology*, 1970, **22**, 37–42. (b)

Kahneman, D. *Attention and effort.* Englewood Cliffs, N.J.: Prentice-Hall, 1973.

Kryter, K. *The effects of noise on man.* New York: Academic Press, 1970.

Woodhead, M. Searching a visual display in intermittent noise. *Journal of Sound and Vibration*, 1964, **1**, 157–161.

Woodhead, M. M. Effect of noise on the distribution of attention. *Journal of Applied Psychology*, 1966, **50**, 296–299.

(Received October 26, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 5, 555-561

THE HARASSED DECISION MAKER:
TIME PRESSURES, DISTRACTIONS, AND THE USE OF EVIDENCE

PETER WRIGHT[1]

Department of Business Administration, University of Illinois at Urbana-Champaign

Dominant simplifying strategies people use in adapting to different information processing environments were investigated. The hypotheses tested were that judges operating under either time pressure or distraction would tend to systematically place greater weight on negative evidence than would their counterparts under less strainful conditions. Six groups of subjects (210 male undergraduates) were presented five pieces of information to assimilate in evaluating cars as purchase options. Three groups operated under varying time pressure conditions, while three groups operated under varying levels of distraction. Data usage models assuming disproportionately heavy weighting of negative evidence provided best-fits to a significantly higher number of subjects in the high time pressure and moderate distraction conditions. Subjects also attended to fewer data dimensions in these conditions.

Perhaps the most pervasive task people face in everyday life is trying to use disparate pieces of information to choose among alternatives: consumer goods, investment portfolios, political candidates, etc. The individual equipped with limited information-handling ability must try to balance his desire to accurately choose options which maximize his ultimate consumption benefits and his equally urgent needs to reduce the cognitive strains of the decision task. Reviews by Slovic (1972) and his associates (Slovic & Lichtenstein, 1971) suggest some of the diverse ways judges may simplify data-handling chores. These reviews also demonstrate that while some structural properties of the available information have been varied in judgment process research, the adaptations judges make under high information load have received surprisingly little empirical attention.

A decision maker's need to simplify should become more urgent when he must operate under a heavy information load. Information load is generally conceived as the amount of data to be processed per unit of time. An increase in information load could therefore result from either increasing the amount of data with which a person must cope or decreasing the time available for processing. Amount of data can itself be increased by either increasing the number of decision-relevant pieces of evidence or by increasing the total amount of information in the immediate environment such that the individual becomes distracted. High information load is perhaps the rule rather than the exception for consumers shopping in noisy, crowded, information-packed retail outlets or managers making decisions under the pressure of deadlines.

Faced with a decision task of challenging complexity, an individual may try to restructure that task into a simpler one. He may use one of several simplifying strategies. For example, he can try to defer an impending decision deadline, physically remove the source of the distraction, or move himself to a more peaceful locale. Even when a person's ability to control time pressure or interference is limited, he may still restructure his task by restricting his attention to certain portions of the incoming data. He may exclude from consideration data about less relevant dimensions, even though he would consider those dimensions sufficiently important to input under less taxing conditions; or he may focus attention on data in certain regions of each dimension. For example, a multiattribute option usually offers outcomes which potentially range from highly desirable to highly undesirable. Pieces of data inform the decision maker of the positive or negative implications of choosing an option. A person may limit his data intake by becoming especially attentive only to data about possible negative (or positive) outcomes.

---

[1] Requests for reprints should be sent to Peter L. Wright, who is now at Graduate School of Business, Stanford University, Stanford, California 94305.

In simplifying as proposed, a decision maker accepts some distortion into his subjectively ideal judgment policy. If he ignores entire dimensions of evidence, he chooses in ignorance of what outcomes to expect on those dimensions. The dimensions he attends to will consequently have relatively greater impact on his judgments than they normally would. If he focuses on negative evidence, he sacrifices awareness of the extent of positive outcomes to be expected, and vice versa. It would be surprising if these two strategies are not used concurrently, that is, the harassed decision maker limiting attention to negative (or positive) evidence on a reduced number of dimensions.

Whether he is more comfortable ignoring positive or negative evidence may depend on the payoff structure of the task. Both Webster (1964) and Canavan (1969) found a negative bias when the decision maker's reward system heavily penalized his false positives while ignoring his successes. In many decision tasks, no such well-defined, externally imposed payoff structure exists. Even so, conditions surrounding the judgment may induce the simplifying judge, who feels he must sacrifice some of the available input, to ignore the positive evidence and insure he is aware of impending negative consequences. One such facilitating condition may be that the option(s) evaluated possess both positive and negative features (Abelson & Kanouse, 1966). Kanouse and Hanson (1971) reviewed several streams of research suggesting that judgments about objects with good and bad attributes are more heavily influenced by negative data. Other research suggests a negative bias might emerge where personal investments (hence, personal losses) are involved and where the judgmental context implies final commitment to the chosen option (Einhorn, 1971; Slovic, 1969). The present study examined one such judgment context, the consumer car-buying decision.

At its extreme, disproportionately heavy weighting of negative evidence amounts to using a conjunctive strategy (Coombs, 1964) with multiple cutoffs separating negative from positive outcomes (and hence, negative from positive evidence). Discovery of a datum suggesting an option does not surpass any cutoff results in outright rejection of that option. Einhorn (1971) has suggested (but not demonstrated) that a conjunctive strategy is an attractive simplifying procedure relative to a linear compensatory strategy. While the rationale offered is plausible, Wright's (1974) results caution that executing a conjunctive strategy may not necessarily be viewed as easy by the decision maker. In any case, a person actually using a compensatory strategy may temporarily adjust his data treatment so that negative data is accentuated without going to the extreme of a strictly noncompensatory conjunctive rule. Unfortunately, when the judge under observation makes errors in translating from input data to output judgments it is difficult to distinguish these two cases.

The hypotheses of this study were that when persons make judgments involving personal investment, a sense of commitment, and possible negative outcomes either under time pressure or when they are distracted, disproportionately heavy weighting of negative evidence occurs. In contrast, similar judgments made under more leisurely conditions reflect no dominant evidence-usage pattern since individual utility functions vary.

METHOD

To test these hypotheses, mathematical models representing an unbiased, a negatively biased, and a positively biased data usage scheme were formulated. All were variations of the general additive compensatory model (Slovic & Lichtenstein, 1971). The general model is:

$$J(X) = b_i X_i + \ldots + b_k X_k, \quad i = 1, 2, \ldots, k \qquad [1]$$

where $J(X)$ is an overall (numerical) judgment of an option; X_i is a (numerical) scale value for that option of the ith dimension; and b_i is the weight given the ith dimension as a whole. When the judgment is recorded on a response scale and the stimuli are expressed as levels on descriptive scales, numbers are assigned to the scale points for entry in this model. The judgment scheme variations contrasted in the present context concerned only the scale values on the right side of Equation 1.

An unbiased model (where negative and positive evidence is equally weighted) was represented by assigning the numbers 1–7 to the seven levels along each descriptive dimension (e.g., "greatly below average" = 1; "greatly above average" = 7). Two negative bias models were used. In one model (Equation 2), the scale values X_i equals 1,...,7 were transformed into log X_i. The effect was that differences between the scale point values increased nonlinearly as the descriptive scales became increasingly negative (e.g., log 1 = 0; log 2 = .301; log 3 = .477, ...; log 6 = .778; log 7 = .845).

$$J(X) = b_i \log (X_i) + \ldots\ldots + b_k \log (X_k),\ i > 0 \qquad [2]$$

Since the descriptive scales used had an average midpoint, negative evidence might have been construed as evidence implying below average features. The second negative bias model transformed only the below average section of the stimulus scales so that the effective scale values were 0, 4, 6.5, 8, 9, 10, 11. The first negative bias model described (Equation 2) is labeled Negative Bias Model 1 and the second Negative Bias Model 2.

Two positive bias models were also used. In Positive Bias Model 1 (Equation 3), differences between the scale values increase nonlinearly as the scales become increasingly positive.

$$J(X) = -b_i \log (a - X_i) - \ldots - b_k \log (a - X_k), \qquad [3]$$

In Equation 3, a is an arbitrary constant set above the highest scale value (7 in this case) so the predicted judgment remains finite. Positive Bias Model 2 transformed only the above average section of the stimulus scales so the effective scale values were 1, 2, 3, 4, 5.5, 8, 12.

The procedure was to fit each of these models to the data of individual subjects making judgments under different time pressure or distraction conditions. If substantially more of the harassed subjects' data usage processes were best matched by the Negative Bias Model 1 or the Negative Bias Model 2, this could be taken as support for the hypotheses.

Judgment Task

Subjects were given descriptions of 30 hypothetical car models. Information on five attributes of each car was given: (*a*) selling price, (*b*) ease of handling, (*c*) cost of maintenance, (*d*) styling, and (*e*) riding comfort. A pilot study had established these attributes as generally salient in evaluating cars. The information was expressed as ratings on 7-point scales with endpoints labeled *greatly below average* and *greatly above average;* the midpoint was labeled *average*. The descriptions were created so that each attribute appeared an approximately equal number of times and so that each car was a mixture of positive and negative attributes. This was important since the models would not discriminate very well where the evidence about an option is fairly homogeneous. Subjects were told the norms implied by the *average* label referred to the class of cars selling for less than $4,000. They were asked to treat the available data as credible and as constituting their own beliefs about the cars.

Subjects judged each car according to the likelihood they would purchase such a car for personal use upon graduation from college. The context was thus one of final preference rather than preliminary screening, and specified an act rather than a general evaluation. In the time pressure study, judgments were recorded on a 4-point scale ranging from extremely high probability to extremely low probability. In the distraction study, the scale used to record judgments was a 7-point bipolar scale with endpoints labeled *likely* and *unlikely*.

Subjects

Subjects were 210 male undergraduates enrolled in a business curriculum and approaching graduation. Each was randomly assigned to one of three time pressure conditions (final cell size of 40) or one of three distraction conditions (final cell size of 30).

Time Pressure Treatments

Three variations in time pressure were created. In the high time pressure condition, subjects were told to make as accurate judgments as possible but were also told that subsequent tasks awaited them. They were asked to proceed as rapidly as possible without sacrificing accuracy. To increase their awareness of time pressure, an assistant recorded elapsed time in 10-second intervals on a visible blackboard. Subjects were asked to record the elapsed time on their booklet when they finished. Subjects in the low time pressure condition were told their only task was to accurately judge the cars. Each was told he would have 40 seconds to consider the information available and should use the entire period. Only when the end of each 40-second interval was signaled by the assistant could he record his judgment and proceed to the next car. The length of a 40-second interval was demonstrated to dramatize that it offered plenty of processing time. In the undefined time pressure condition, instructions were similar to the low time pressure condition, but no mandatory deliberation period was imposed. Subjects were told to proceed at whatever pace suited them. After completing the judgment task, subjects were asked, "How much time pressure did you feel while making your judgments?" They responded on a 5-point scale with endpoints labeled *very much pressure* and *very little pressure*.

Distraction Treatments

Distraction treatments were not crossed with time pressure treatments. Three levels of distraction were created. In all three, subjects were given an introduction similar to undefined time pressure condition and were also forewarned that some noise would accompany their task to simulate a natural decision environment. In the high distraction condition, a taped excerpt from a radio talk show (question and answer format, including commercials) was played at a moderately high volume through the remainder of the task. In the moderate distraction condition,

TABLE 1

FREQUENCY OF BEST-FITS FOR DATA USAGE MODELS IN TIME PRESSURE AND DISTRACTION TREATMENTS

Data usage model	Time pressure		
	High	Low	Undefined
Unbiased	5	14	15
Negative bias (1)	26	11	13
Positive bias (1)	9	15	12
	Distraction		
	High	Moderate	Low
Unbiased	9	5	11
Negative bias (1)	9	18	9
Positive bias (1)	12	7	10

the same tape was played at low volume. In the low distraction condition, taped background music from a frequency modulation (FM) station was played. In all conditions subjects were assured their primary task was to accurately evaluate the cars. After the judgment task, each was asked, "How distracting did you find the noise from the tape recording while making your judgments?" They responded on a 5-point scale with endpoints labeled *very distracting* and *not very distracting*. In addition, they were asked to describe any methods they used to handle the distractions.

RESULTS

Mean time-per-judgment recorded by subjects in the high time pressure condition was 12.2 seconds compared with the standard 40 seconds in the low time pressure group. Exact timekeeping for subjects in the undefined time pressure condition was difficult, but experimenters' estimates showed an average total time of 10 minutes, or about 20 seconds per judgment. Time *used* is not the optimal measure of perceived time pressure, however. Mean ratings of perceived time pressure were 4.67, 2.70, and 2.12 for the high time pressure, low time pressure, and undefined time pressure treatments, respectively. Analysis of variance revealed a significant difference between the three groups ($F = 8.05$, $df = 2/118$, $p < .01$). High time pressure subjects felt more time pressure than did either of the other groups, which did not differ significantly.

Subjects in the distraction conditions were asked how distracting they found the extraneous noise accompanying their task. Means were 4.30, 3.19, and 1.35 for high distraction, moderate distraction, and low distraction treatments, respectively ($F = 6.74$, $df = 2/88$, $p < .01$). Neuman-Keuls analysis showed each treatment differed significantly from each of the others.

For each subject, multiple correlations were computed between his actual judgments and those predicted by the unbiased model, the two negative bias models, and the two positive bias models. For the last four, the appropriate transformations were made before the scale values were entered into the regression. Two values for a (8 and 50) were used in Equation 3 to see what differences would occur in the results. The effect was minimal, and the reported results are for $a = 8$. Analysis at the level of the individual was most relevant to the question of simplifying strategies. For each subject, the model yielding the highest multiple correlation was noted. The frequency with which each model gave the highest multiple correlation is shown in Table 1.

In computing these frequencies, the negative bias models were not contrasted against each other but were used alternately. The same was true for the positive bias models. Negative Bias Model 1 turned out to be virtually interchangeable with Negative Bias Model 2. Similarly, Positive Bias Model 1 was interchangeable with Positive Bias Model 2. Substitutions yielded only two reclassifications. Consequently, the frequencies shown in Table 1 are for Positive Bias Model 1 and Negative Bias Model 1. Similar analyses for frequencies of the highest multiple correlations produced by all other model combinations gave similar results.

The hypothesis was that operating under pronounced time pressure or distraction induces a general tendency among subjects to rely heavily on negative evidence. Examining first the time pressure effect, no systematic pattern in the weighting tactics of low time pressure or undefined time pressure subjects is apparent. All three models provide optimal fits for about the same number of subjects.

However, approximately two thirds of the subjects operating under high time pressure were best fit by the model assuming heavy weighting of negative data. An overall chi-square test gave a value of 14.62 ($df = 4$, $p < .01$). Comparing the high time pressure subjects with the collapsed samples from the other two conditions gave a chi-square value of 13.83 ($df = 2$, $p < .001$). Subjects forced to assimilate multiple cues under time pressure did differ from their counterparts handling the same information under less pressure. Thus, the hypothesis that subjects under pronounced time pressure rely heavily on negative data was supported.

In the distraction study, the emerging patterns were somewhat different. Again, the least strainful condition (low distraction) produced no evidence that subjects displayed anything but personal idiosyncracies in the way they weighted data. Greater frequency of negative bias did occur when a moderate level of distracting noise surrounded the task. Sixty percent of the subjects in the moderate distraction treatment were best fit by Negative Bias Model 1. Yet this pattern did not repeat itself when distractions increased (high distraction). Overall chi-square analysis gave a value of 8.62 ($df = 4$, $p < .10$). Comparing the moderate distraction group against the collapsed low distraction and high distraction groups gave a chi-square value of 7.30 ($df = 2$, $p < .05$). This analysis offers tentative support for the hypothesized dependence on negative data when distractions place a strain on attention.

If limiting the nature and amount of the data used is a preferred tactic for handling high information load, the number of separate dimensions subjects consulted in making their judgments might reflect this. Consequently, the number of dimensions with statistically significant ($p < .05$) regression coefficients was calculated for each subject. This gave an estimate of how many factors had been systematically used by the processor to make the final judgment. The maximum was five. Mean number of significant dimensions per subject was 1.50, 2.33, and 2.08 for the high time pressure, low time pressure, and undefined time pressure conditions, respectively. One-way analysis of variance (ANOVA) indicated significant differences between the three conditions ($F = 4.45$, $df = 1/118$, $p < .05$). Neuman-Keuls analysis showed the high time pressure group used significantly fewer dimensions than did the other two groups, which did not differ. In the distraction study, the mean number of dimensions per subject was .98, 1.63, and 2.15 for the high distraction, moderate distraction, and low distraction groups, respectively. Similar analyses showed the high distraction subjects used fewer dimensions than did either the moderate distraction or low distraction subjects, and the moderate distraction group used fewer than did the low distraction group ($F = 6.32$, $df = 1/88$, $p < .05$).

TABLE 2

MEAN MULTIPLE CORRELATIONS FOR TIME PRESSURE AND DISTRACTION TREATMENTS

Data usage model	Time pressure			
	High	Low	Undefined	
Unbiased	.56	.70	.69	*.65*
Negative bias (1)	.62	.72	.70	*.68*
Positive bias (1)	.54	.68	.72	*.65*
	.57	*.70*	*.70*	
	Distraction			
	High	Moderate	Low	
Unbiased	.50	.63	.72	*.61*
Negative bias (1)	.48	.68	.70	*.62*
Positive bias (1)	.47	.62	.73	*.61*
	.48	*.65*	*.71*	

Just how adequate were any of these models in describing the strategies subjects used? Frequent best-fits for Negative Bias Model 1 where predicted would be less meaningful if multiple correlations were low. Mean multiple correlations for the models are shown in Table 2. After transforming the multiple correlations via Fisher's z transformation, separate 3×3 ANOVAs were run for the time pressure and distraction studies. In the time pressure study, the main effects of time pressure ($F = 46.34$, $df = 2/118$, $p < .001$) and model ($F = 6.86$, $df = 2/236$, $p < .05$) were significant, as was the interaction ($F = 36.23$, $df = 4/236$, $p < .001$). The time pressure effect, due to the somewhat lower correlations in the high time pressure condition, was anticipated by the preceding analysis of number of significant dimensions per subject. The interaction, also expected from preceding individual level analyses, was due to higher correlations for Negative Bias Model 1 in the high time pressure condition versus the comparably high correlations for all three models in the low time pressure and undefined time pressure conditions.

The repeated-measures ANOVA for the distraction study yielded a significant main effect for distraction ($F = 67.62$, $df = 2/88$, $p < .001$) and a significant interaction effect ($F = 12.88$, $df = 4/176$, $p < .001$). The model's effect was not significant ($F = 2.32$, $df = 2/176$, ns). The expected interaction was due to the relatively higher correlation for Negative Bias Model 1 in the moderate distraction condition versus comparable correlations in the other distraction conditions. The relatively poor-fits generally found in the high distraction condition suggest the manipulation may have disrupted the subjects' processing so much they became erratic. Hence the negative bias found among moderate distraction subjects failed to hold for the overstrained high distraction subjects. Such an interpretation is partially supported by subjects' posttask descriptions of how they coped with the distractions. High distraction subjects tended to report frustration and an awareness of inconsistency more often than did moderate distraction subjects.

As one final test of how well a negative bias model did in explaining the adjustments of subjects in the high time pressure and moderate distraction conditions, the maximum multiple correlation for subjects in these two conditions was compared to the median maximum multiple correlation for the two conditions. In the high time pressure condition, 17 of the 26 subjects best fit by Negative Bias Model 1 fell above the median maximum multiple correlation for that condition (.635). The corresponding figures for the unbiased and positive bias models were 1 of 5 and 3 of 9, respectively. In the moderate distraction condition, 10 of the 18 subjects best fit by Negative Bias Model 1 fell above the sample median (.673). Corresponding figures for the unbiased and positive bias models were 3 of 5 and 3 of 7, respectively.

DISCUSSION

The proportionately greater frequency of best-fits afforded by the negative bias model in the high time pressure and moderate distraction conditions appears to support the original hypotheses. A tendency for people to accentuate negative evidence when the environment discourages leisurely processing may be indicated. A complementary tendency to use fewer attributes in the same circumstance is also indicated. The harassed decision maker is pictured as becoming extremely alert to discrediting evidence on a few salient dimensions.

The perspective taken in the present study on how people use evidence in evaluating alternatives echoes that of Shepard (1964) and Tversky (1972) in emphasizing a person's immediate "state of mind" as the major determinant of the weights applied. Repeated findings of task effects on data weighting policies and of intransitivity in choices seem to argue against traditional ideas of a stable utility function dictating a weighting scheme which consistently molds a decision maker's evaluation policy across situations. The weight a person gives to some aspect of the decision problem is in part a function of stable goals and in part a function of immediately salient subgoals. The latter may dominate the former under certain circumstances, and vice versa. The questions of interest then become identifying the conditions that lead a judge to deviate from his optimal rational strategy (if such a notion is meaningful) and discovering whatever stability exists in the manner in which he deviates under those conditions. Situational factors related to information overload, two of which were examined in this study, seem a promising starting point.

A limitation of fitting mathematical models to input–output data is the remaining ambiguity about the underlying processes. Each of the models used in this study was literally an additive, compensatory model. The conservative interpretation of the results must therefore be in terms of differential weighting of negative evidence by compensatory data processors. However, both negative bias models used in this study may be viewed as approximations to a noncompensatory conjunctive strategy. For example, Einhorn (1971) has proposed that if Equation 2 incorporates a log transform of the response scale values on the left side of the equation, it may be treated as a conjunctive model. Since Negative Bias Model 2 bends the scale values even more sharply than does Negative Bias Model 1, it might also be seen as a reasonable approximation to a multiple cutoff strategy. Some may therefore be willing to interpret the results as indicating more frequent use of multiple cutoff strategies under difficult conditions. Several researchers have, however, cautioned against treating math models like these as close enough approximations to a conjunctive model to warrant such an interpretation (e.g., Birnbaum, 1973; Goldberg, 1971). Clearly, more rigorous tests, relying more on introspective reports from subjects than this study did, are necessary to sort out the precise interpretation.

If these subjects had had considerable experience integrating evidence about cars under difficult conditions, they might have developed individualized shortcuts idiosyncratic to the topic of cars. An interesting question for future research is whether people who consistently operate in noisy, pressurized environments continue to accentuate negative evidence or develop other shortcuts. Such people might have had the benefit of feedback which the subjects in this study did not. It is not clear whether the types of feedback usually available to consumers or managers would sensitize them to the deviations they introduce in trying to simplify. At a minimum they would need feedback under ideal conditions for insight into their preferred judgment policy, plus feedback about the outcomes of judgments made when harassed, soon enough after making that judgment, to recall the conditions surrounding it. It would seem much experience is necessary for a person to realize the locus of incorrect judgments lies in his simplifying adaptations, not in his preferred judgment policy. It seems it is time that judgmental research became more concerned with the task conditions systematically affecting the use of different judgmental strategies rather than general model building without concern for the task environment.

REFERENCES

Abelson, R. P., & Kanouse, D. E. Subjective acceptance of verbal generalizations. In S. Feldman (Ed.), *Cognitive consistency*. New York: Academic Press, 1966.

Birnbaum, M. H. The devil rides again: Correlation as an index of fit. *Psychological Bulletin,* 1973, **70**, 239–242.

Canavan, D. *The development of individual differences in the perception of value and risk-taking style.* Unpublished doctoral dissertation, Columbia University, 1969.

Coombs, C. H. *A theory of data.* New York: Wiley, 1964.

Einhorn, H. J. Use of nonlinear, noncompensatory models as a function of task and amount of information. *Organizational Behavior and Human Performance,* 1971, **6**, 1–27.

Goldberg, L. R. Five models of clinical judgment: An empirical comparison between linear and nonlinear representations of the human inference process. *Organizational Behavior and Human Performance,* 1971, **6**, 458–479.

Kanouse, D. E., & Hanson, L. R. Negativity in evaluations. In E. E. Jones et al., *Attribution: Perceiving the causes of behavior.* Morristown, N.J.: General Learning Press, 1971.

Shepard, R. N. On subjectively optimum selection among multiattribute alternatives. In M. W. Shelly II & G. L. Bryan (Eds.), *Human Judgments and Optimality*. New York: Wiley, 1964.

Slovic, P. Differential effects of real versus hypothetical payoffs on choices among gambles. *Journal of Experimental Psychology,* 1969, **80**, 434–437.

Slovic, P. From Shakespeare to Simon: Speculations—and some evidence—about man's ability to process information. *Oregon Research Institute Monograph,* 1972, **12**, No. 2.

Slovic, P., & Lichtenstein, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance,* 1971, **6**, 649–744.

Tversky, A. Elimination by aspects: A theory of choice. *Psychological Review,* 1972, **79**, 281–299.

Webster, E. C. *Decision making in the employment interview.* Montreal: Industrial Relations Centre, McGill University, 1964.

Wright, P. The choice of a choice strategy: Simplification vs. optimization. Faculty Working Paper, Bureau of Economic and Business Research, University of Illinois, 1974.

(Received July 30, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 5, 562–571

EXPERT JUDGMENT:

SOME NECESSARY CONDITIONS AND AN EXAMPLE[1]

HILLEL J. EINHORN[2]

Graduate School of Business, University of Chicago

Expertise is examined within situations where judges deal with multidimensional information. Psychometric criteria were advocated as being indicative of expert judgment: (*a*) Experts should tend to cluster variables in the same way when identifying and organizing cues. (*b*) Expert judgment should be highly reliable (intrajudge reliability), show both convergent and discriminant validity, and be relatively free of judgmental bias when measuring cues. (*c*) Experts should weight and combine information in similar ways. These criteria were applied to data obtained when three medical pathologists rated the amount of histological signs in biopsy slides. The results showed that these experts generally met the criteria, although they did not seem to weigh information similarly. Some limitations of defining expertise in terms of the above criteria are discussed.

In an age of increasing specialization, it is quite likely that greater numbers of people will have to rely on expert judgment. It is therefore an important matter to discuss some of the conditions that produce "expertise." Is an expert simply one who is very skillful with training and knowledge in some specialized field or are there perhaps more objective criteria that can be used? The purpose of this study is to discuss and illustrate some necessary, if not sufficient, conditions for defining expertise within a given situation. The type of situation to be considered is one in which several judges have to deal with multidimensional information. This information has to be measured and combined into a decision or global judgment. This is a very general task, specific examples of which might include: (*a*) radiologists viewing an X ray and coming to a diagnosis; (*b*) businessmen making a sales forecast for a new product; (*c*) art critics viewing a painting and judging its quality, etc.

The expert must *identify* information or cues from the multidimensional stimulus he encounters. These cues are diagnostic (i.e., contain information) about the final decision or judgment. The expert's ability to identify cues can be seen as a problem of extracting weak signals from a background of noise (e.g., Green & Swets, 1966). While each cue is related to the final categorization, it also serves the function of leading the expert to other cues that will also be of importance in the determination of the final categorization. For example, consider the cue of "product safety". While this cue can be related to how much sales we can expect, it is also indicative of other cues in the product (its price may be higher due to safety features). When one considers that the expert may deal with many cases in the course of time, it is not unreasonable to assume that he builds up expectations regarding both what variables typically covary together and the strength of this covariation.

Consider that the expert must measure two cues that he considers to be highly related, for example, the degree of workmanship and the quality of material in a particular product. If he believes that these cues covary together, it is quite likely that his measurement of one cue will affect his measurement of the other. This is similar to what has been termed a *logical error* (Guilford, 1954); that is, traits that logically go together are rated in similar ways. The relationship different experts expect between cues may vary due to their different experience and training. Furthermore, this is not an error but reflects each expert's way of *organizing information into clusters*

---

[1] This research was supported by a grant from the Spencer Foundation.

[2] Requests for reprints should be sent to Hillel J. Einhorn, Graduate School of Business, University of Chicago, Chicago, Illinois 60637.

or dimensions.[3] From a psychological point of view, this clustering of information serves to reduce the dimensionality of the information one has to process; that is, instead of having to process many cues, one only has to deal with a smaller number of general factors.

While cues are being identified and clustered together, the expert is *measuring the amount* of the cue. The businessman may say that a product is "tasty", but *how* "tasty"? In many types of situations the expert will have to serve as a measuring instrument for cues that cannot be measured in more objective ways (cf. Sawyer, 1966).

There are three issues that are important in discussing the measurement of cues: (*a*) intrajudge reliability, (*b*) construct validity (Cronbach & Meehl, 1955), and (*c*) judgmental bias. With regard to intrajudge reliability, it should be obvious that unless the expert can reproduce his measurements of the cues, there is little more that can be said in defense of his expertise. Therefore, high intrajudge reliability is a necessity for expertise. With regard to construct validity, it should be the case that the cues being measured should have status as explanatory concepts. Operationally, we would expect that the cues show both convergent and discriminant validity when they are considered as traits (judges are considered methods) in a multitrait–multimethod matrix (Campbell & Fiske, 1959). Finally, with regard to judgmental bias, we should expect that expert judges will not show such well-known biases as *leniency, halo,* etc. (Guilford, 1954; Ghiselli & Brown, 1955). All three of the above criteria should be considered in evaluating expertise.

After cues are identified, measured, and clustered, the important cognitive work of *weighting and combining* to form a global evaluation follows. This part of the judgment task has received the greatest amount of attention (Slovic & Lichtenstein, 1971). The results of a large number of studies have shown that judges are quite poor at weighting and combining information accurately, that is, in accord with the objective validity of the information. However, objective validity implies that the criteria used as the bases for comparing the global evaluation were themselves relevant to the ultimate criteria in question (Thorndike, 1949). In many cases, availability of criteria, not relevance, was used to define the objective measure. It is therefore possible that global judgments have more validity than it has been possible to demonstrate thus far.

Since questions of criterion relevance cannot be fully answered, and since some situations contain no criteria, is there some other way of evaluating global judgments? Let us take the same line of inquiry that was used in the measurement of traits; that is, Is the global evaluation an operational concept in the sense of possessing convergent validity? While it would seem important that experts agree on their evaluations (as a condition for expertise), the issue is considerably more complicated. The concept of agreement can itself be thought of in two ways: (*a*) agreement "in fact" and (*b*) agreement "in principle". The former refers to actual agreement of evaluation, whereas the latter refers to agreement with respect to weighting and combining policy, that is, *how* the global evaluation is to be formed once the inputs are specified. Although one may have agreement in principle, there may be disagreement with respect to the coding of input (note that coding of input is also dependent on the perceived covariation of the cues). Therefore, agreement in fact will be a function of weighting and combining similarity *and* similarity in coding cues. On the other hand, it should not be assumed that agreement in principle is easy to achieve. This is due to the fact that learning probabilistic relationships in the natural environment may be quite difficult. This implies that there is great room for individual differences in the evaluation of the diagnosticity of cues and subsequently, large differences in the weights assigned to those cues. Furthermore, even when training is rigorous and the conditions for learning optimal, similar weighting policies may reflect training in a particular school of thought. It seems that policies would be very similar within but not between competing schools (a practical application of this is Naylor & Wherry's, 1965, technique for grouping judges on the basis of similarity of their weighting policies).

---

[3] It has been noted that in situations where an objective measure of the relationship between cues is given, subjects find it very difficult to learn them if they are different from their preconceived ideas of what "go together" (Chapman & Chapman, 1969).

Although the difficulty of achieving either type of agreement is great, it should be the case that there is at least a common core of knowledge which is germane to a particular area of inquiry. Furthermore, expertise consists, to some extent, of knowing that common core. Therefore, as a practical guide it is proposed that there be at least some agreement, both in fact and in principle, for the global evaluation.[4]

With the discussion of the factors involved in expert judgment, there is a need to illustrate, in a concrete manner, how the various criteria discussed in this study can be used in evaluation of real experts. The purpose of the present study is to make explicit the processes assumed to be present in the preceding discussion. The details of the study to be described appear in Coppleson, Factor, Strum, Graff, and Rappaport (1970) and Einhorn (1972).

METHOD

Subjects

The judges we considered were three medical pathologists. This specialty requires considerable training and is of special importance since the pathologists' report is often the ultimate criterion that is used for the diagnosis of various diseases.

Procedure

The three pathologists independently viewed 193 biopsy slides taken from patients having Hodgkin's disease, a cancer of the lymph system. The biopsies were taken when the patient first entered the hospital. The only information given to the pathologists was the biopsy slide. All the slides used in this study were taken from patients who had died and from whom complete data were available with respect to the ratings to be described.

For each slide, the pathologist had to give his judgment of the amount of nine histological characteristics that were chosen, a priori, to be important. The histological signs were as follows: benign histiocytes, malignant histiocytes, eosinophils, necrosis, plasma cells, neutrophils, lymphocytes, Sternberg-Reed distribution, and Sternberg-Reed cells. For each of these signs, with the exception of Sternberg-Reed distribution (which was measured on a 2-point scale), a 6-point scale was used, which was developed by the pathologists.

In addition to the judgments of the amounts of the histological characteristics, a global judgment concerning the classification of the disease in terms of severity was also made. This judgment was measured on a 9-point scale (higher values representing greater severity). The global judgment was made after the judgments of the histological signs. Finally, 26 of the slides were repeated twice each so that estimates of test–retest reliability (the ability to reproduce the judgments) could be obtained.

Although the judges picked out the signs they expected to measure, one would expect that experts with experience in this task would have little difficulty in recognition of attributes, whereas the judgment of amount would be of most concern. However, definitions of each sign were not given so that any differences the judges had in this regard would be revealed in the later analyses. Of special importance is the fact that the multidimensional stimulus was viewed intact, so that interactions, context effects, etc. could play whatever role they normally did in the task.

RESULTS

The first question addressed was whether the ratings or cue measurements were reliable; that is, Could the expert, when viewing exactly the same slide on two occasions, give the same response? Mean intrajudge reliability for the signs ranged from .19 to .93 (over the three judges), while the mean global judgment reliability was .63 (over the three judges). The mean reliability for the three judges (over signs) ranged from .64 to .74. Judge 2 had a slightly higher reliability for the signs, but his reliability on the global judgment (.46) was considerably lower than that of the other judges (.69, .71). The average reliability over both judges, signs and global judgments was .69.

Convergent and discriminant validity were examined for evidence of construct validity. Convergent validity refers to the fact that different methods (judges) should converge in measuring the same trait (sign); the correlation between judges for the same sign should be greater than zero. Discriminant validity, on the other hand, is defined by three conditions: (*a*) The correlation between the same traits (signs) for different judges should be higher than the correlation between different traits as measured by different judges. (*b*) The correlations between the same signs as measured by different judges should be higher than the correlations between different signs as measured by the same judge. (c) The pattern of correlations should be the same in the submatrices of the multitrait–multimethod matrix. This refers to the ordering in both sign and magnitude of the correlations in the heterotrait–heteromethod matrices.[5]

---

[4] In cases where experts are from competing schools, we might expect high negative correlations.

The multitrait–multimethod (multisign–multijudge) matrix is shown in Table 1.

Looking first at convergent validity, 29 of the 30 correlations between judges' measures of signs were significant at $p < .01$. Furthermore, the average of these correlations was .56—evidence of convergent validity. It should be noted that convergent validity, in situations where a single judge is defined as a method, is the same as *inter*judge reliability, or agreement. The amount of agreement was considerably different for the various signs (.33 to .80). Notice should be given to the fact that the greatest amount of disagreement was with respect to the global judgment (.27). This finding is in accord with earlier remarks about the global judgment being a composite variable. As far as the mean agreement for the three pair of judges, over all signs, Judges 1 and 2 agreed with each other most (.64), with Judges 1 and 3 next (.52), and Judges 2 and 3 agreeing least (.50).

In analyzing discriminant validity of 540 comparisons, there were only 20 violations (or 4%) of the requirement that the correlation between the same signs for different judges be higher than the correlation between different signs as measured by different judges. Additionally, there were only 32 violations of the 540 comparisons (or 6%) of the requirement of discriminant validity that correlations between the same signs as measured by different judges be higher than the correlations between different signs as measured by the same judge. Given that this requirement is quite stringent and the percentage of violations small, it seems fair to conclude that this condition was also met. Examination of the submatrices further showed that the pattern of correlations was the same in the submatrices, indicating that the third condition for discriminant validity had been met. Therefore, it seems fair to conclude that this data showed both convergent and discriminant validity.

Guilford (1954) presented an analysis of variance approach for dealing with judgmental biases found in the use of rating scales. While the situation that this approach was originally developed for involved raters rating employees on various traits, the general idea is applicable to any situation where judges rate objects on a number of dimensions (Stanley, 1961, also discusses this basic design). Although Guilford's procedure is applicable to the present data, there is a slight difference. In Guilford's example (pp. 280–288), each rater rates each ratee on each attribute only once. In the present study, using the 26 repeat slides, we have a Judges × Slides × Signs design with two replications. The last two factors are repeated and therefore an appropriate way to analyze this data is given by Winer (1962, pp. 319–337). This analysis of variance is presented in Table 2.

Note first that the main effect for judges is not significant. Since this effect represents the judges' overall tendency to rate the amount of the signs, this indicates a lack of a *leniency* error. (Leniency in this case refers to a tendency to overvalue or undervalue the slides in general). The lack of an effect in the present context is more in accord with what one would expect from expert judges. The two other main effects simply indicate that different slides and different signs have varying means. What is of major concern are the interactions involving the judges' factor. These involve the differential effect of judges on slides (A × B), on signs (A × C) and on Slides × Signs (A × B × C). These interactions should be nonsignificant; that is, differential effects due to judges are not consistent with the fact that the shared core of expertise is influencing their measurement. Consider the Judges × Slides (A × B) interac-

---

[5] These conditions can be expressed formally. Let, x_{ijs} = rating on ith slide by jth judge for sth sign. Let $i \neq i'$, $j \neq j'$, $s \neq s'$. Convergent validity then refers to $\rho_{x_{.js}, x_{.j's}} > 0$ (correlating over i). The two conditions for discriminant validity are:

(*a*) $\rho_{x_{.js}, x_{.j's}} > |\rho_{x_{.js}, x_{.j's'}}|$; (*b*) $\rho_{x_{.js}, x_{.j's}} > |\rho_{x_{.js}, x_{.js'}}|$.

Absolute value operators are used, since it is the magnitude rather than the sign of the correlation that is important.

TABLE

MULTITRAIT (SIGNS)-

	Judge 1										Judge 2			
	Lukes	BH	MH	Eos.	Nec.	PC	Neut.	Lymph.	S-R D.	S-R C.	Lukes	BH	MH	Eos.
Judge 1														
Lukes	(690)													
BH	−044	(430)												
MH	302	−061	(820)											
Eos.	217	−116	138	(860)										
Nec.	058	100	370	134	(840)									
PC	126	−035	061	104	105	(470)								
Neut.	212	114	193	294	302	189	(490)							
Lymph.	−310	−212	−447	−420	−288	−159	−422	(850)						
S-R D.	280	−094	599	172	160	007	184	−339	(220)					
S-R C.	306	−131	449	027	191	029	129	−139	461	(590)				
Judge 2														
Lukes	400	016	135	−091	116	038	−086	−021	160	275	(460)			
BH	−041	446	049	−108	190	−012	027	−231	−017	−083	051	[illegible]		
MH	241	−096	634	071	144	060	082	−318	484	402	179	[illegible]	(860)	
Eos.	109	−037	114	824	099	069	266	−387	119	−048	−113	[illegible]	015	(900)
Nec.	121	128	382	142	858	069	301	−320	230	195	108	[illegible]	180	129
PC	013	−098	−047	132	069	600	161	−157	−115	−100	−051	[illegible]	−086	131
Neut.	239	−007	235	294	368	217	655	−471	270	153	033	[illegible]	139	339
Lymph.	−217	−133	−356	−348	−316	−088	−424	686	−264	−120	−062	[illegible]	−282	−402
S-R D.	160	−078	285	−037	056	072	056	−118	489	266	169	[illegible]	451	−033
S-R C.	156	−093	315	−018	080	044	031	−115	332	596	169	[illegible]	511	−074
Judge 3														
Lukes	209	069	233	090	170	093	101	−225	141	229	199	[illegible]	232	054
BH	−037	805	−093	−072	−129	020	−103	−104	072	−042	−040	[illegible]	−047	−020
MH	222	053	500	084	354	111	214	−351	365	412	184	[illegible]	503	−005
Eos.	190	−058	075	776	094	050	322	−409	169	049	−035	−017	027	783
Nec.	118	031	282	188	572	154	479	−442	194	140	029	076	097	158
PC	139	−008	−004	086	094	420	189	−180	−061	011	−015	122	−019	057
Neut.	265	004	195	277	362	159	674	−516	206	100	016	070	080	261
Lymph.	−188	−239	−324	−245	−318	−115	−393	614	−215	−131	−052	−182	−195	−220
S-R D.	255	−080	252	319	162	086	272	−260	365	295	075	−043	179	251
S-R C.	275	−087	298	120	221	049	215	−173	386	558	267	−151	275	022

Note. Correlations are based on $n = 193$. Validity diagonals are circled. Correlations in parentheses are intrajudge reliabilities ($n = 26$). Lukes = judgment of severity; BH = benign histiocytes; MH = malignant histiocytes; Eos. = eosinophils; Nec. = necrosis; PC = plasma cells; Neut. = neutrophils; Lymph. = lymphocytes; S-R D. = Sternberg-Reed distribution; S-R C. = Sternberg-Reed cells.

tion. This interaction involves the particular judge's tendency to overvalue or undervalue particular slides. This can be conceptualized as a *halo* effect; that is, a general tendency dominates the ratings of the particular attributes. In this situation, a halo effect would mean that the pathologists first get a general impression of severity and then work at rating the signs in accord with this general impression. However, this interaction is not significant indicating the absence of any halo. Both the Judges × Signs (A × C) and the Judges × Signs × Slides (A × B × C) are significant, although the magnitude of these interactions is quite small. The former interaction indicates that the judges differentially rate the signs; that is, some judges *see* more or less of some signs than others. The three-way interaction indicates that judges differentially rate signs depending on what slide the sign appears in. This is a type of context effect that is specific to the judges. However, it should be stressed that the significant interactions are all quite small. For these particular experts then, there seems to be only slight bias in their judgmental ratings.

One final analysis is necessary to complete the results on bias. Because of the qualitative difference between the global judgment and the individual cues, a separate one-way analy-

1

MULTIMETHOD (JUDGES) MATRIX

2						Judge 3									
Nec.	PC	Neut.	Lymph.	S-R D.	S-R C.	Lukes	BH	MH	Eos.	Nec.	PC	Neut.	Lymph.	S-R D.	S-R C.
(840)															
053	(530)														
374	191	(653)													
−305	−127	−536	(760)												
100	−109	076	−088	(170)											
043	−065	045	−072	386	(830)										
168	049	054	−197	049	193	(710)									
−137	032	−126	−031	130	056	−159	(400)								
367	−009	224	−279	291	323	335	−119	(720)							
094	082	315	−406	017	009	119	−006	072	(950)						
561	106	510	−393	115	142	086	−060	309	191	(550)					
032	527	191	−170	−071	048	175	100	028	136	165	(570)				
341	102	652	−446	136	105	113	−016	186	343	672	162	(550)			
−331	−046	−385	629	−126	−102	−229	−089	−436	−331	−443	−158	−455	(730)		
141	045	327	−231	124	201	056	011	250	306	351	002	345	−164	(160)	
245	−129	207	−169	278	423	343	−080	427	133	258	−043	276	−302	407	(590)

sis of variance (repeated measures on each slide for the three judges) was performed on the global judgments. The results showed a significant effect with Judges 1 and 2 being most different ($\bar{Y}_1 = 2.98$, $\bar{Y}_2 = 4.61$). Therefore, Judge 2 saw more severity generally than did Judge 1, although both judges did *not* see different amounts of the signs. As has been pointed out before, this is consistent with the remarks made concerning the difficulty of achieving agreement on global variables.

The issue of identifying what the relevant cues were in this task was simply decided by having experts agree on a set of attributes. While this obviously is useful from an experimental point of view, its external validity is quite limited. To investigate the clustering of relevant cues, the correlations in the monojudge submatrices of the multitrait–multimethod matrix could be examined. A simplifying method for comparing the correlational structure of the matrices is to compare their factor structure. In order to achieve this, an orthogonal factor analysis with varimax rotation was performed on the correlation matrices of the cues, one for each of the three judges. Four factors were extracted, accounting for 70% of the variance of each matrix. These results are shown in Table 3.

TABLE 2

ANALYSIS OF VARIANCE FOR JUDGES × SLIDES × SIGNS

Source	*SS*	*df*	*MS*	*F*
Judges (A)	2.38	2	1.19	<1
Subjects × Within	3.88	3	1.29	
Slides (B)	160.70	25	6.43	14.61**
A × B	21.10	50	.44	1.47
B × Subjects Within	22.75	75	.30	
Signs (C)	975.63	8	187.80	257.26**
A × C	26.37	16	1.65	2.26*
C × Subjects Within	17.62	24	.73	
B × C	976.08	200	4.88	13.94**
A × B × C	159.15	400	.40	1.14*
B × C × Subjects Within	208.93	600	.35	

Note. All factors are assumed fixed.
* $p < .05$.
** $p < .001$.

The rotation accomplished a good degree of simple structure as can be seen. Factor loadings of .5 or greater (in either direction) are circled. Judges 1 and 2 seem to have a much more similar structure than Judge 3. In order to adequately compare the factorial structure of the judges, factor scores were obtained for each judge. This was accomplished via the least squares procedures outlined in Thurstone (1947). The factor score, for each judge, represents that judge's score on the underlying construct represented by that factor. In order to compare factor matrix similarity, Nunnally (1967) has stated that factor scores be correlated across the different analyses. Since there are three judges with four factors, the results from an application of this procedure can be conceptualized within a multitrait–multimethod framework, that is, multifactors–multijudges.[6] Table 4 shows the Judges × Factors matrix of correlations.

The correlations within the single judges' matrices are all zero due to the orthogonal factor analysis. The validity diagonals are circled. It can be seen that for Judges 1 and 2, the factorial structure is quite similar (the average of these values is .62). In addition, this matrix shows high discriminant validity (in the sense of the first condition). However, note that both comparisons of the matrices involving Judge 1–Judge 3, and Judge 2–Judge 3 only show convergent validity on the first two factors. For Factors 3 and 4 there is no convergent validity (and therefore no discriminant validity). The fact that

---

[6] Let f_{ijr} = factor score on *i*th slide for *j*th judge on the *r*th factor. Then, convergent validity of factor structure is defined as, $\rho_{f.jr,f.j'r} > 0$. Since the factors are orthogonal, $\rho_{f.jr,f.jr'} = 0$. Condition (*b*) for discriminant validity cannot be tested. Condition (*a*), however, becomes, $\rho_{f.jr,f.j'r} > |\rho_{f.jr,f.j'r'}|$.

TABLE 3

FACTOR MATRICES FOR THREE JUDGES

Signs	Judge 1				Judge 2				Judge 3			
	Factor				Factor				Factor			
	I	II	III	IV	I	II	III	IV	I	II	III	IV
Benign histiocyte	−.142	−.080	.906	−.086	−.112	.153	.901	.002	−.035	.018	.959	.099
Malignant histiocyte	.801	.191	.143	.090	.807	.208	.011	−.064	.856	−.055	−.095	.124
Eosinophils	−.001	.857	−.133	−.050	−.184	.706	−.387	−.008	−.075	.735	.047	.165
Necrosis	.369	.166	.402	.422	.147	.549	.245	−.042	.370	.466	−.232	.502
Plasma cells	−.056	.058	.065	.875	−.087	.121	−.001	.977	−.103	−.008	.129	.738
Neutrophils	.146	.604	−.042	.414	.074	.773	.040	.178	.236	.631	−.159	.480
Lymphocytes	−.309	−.694	−.414	−.113	−.135	−.801	−.174	−.055	−.559	−.232	−.119	−.522
Sternberg-Reed distribution	.778	.222	−.052	−.110	.738	.046	.025	−.105	.332	.732	.071	−.225
Sternberg-Reed cells	.796	−.100	−.140	.081	.808	−.039	−.104	.063	.733	.296	−.007	−.191

Note. Factor loadings of .5 or greater (in either direction) are circled.

TABLE 4

INTERCORRELATION OF FACTOR SCORES FOR THREE JUDGES

Factor		Judge 1				Judge 2				Judge 3			
		1	2	3	4	1	2	3	4	1	2	3	4
Judge 1	1	1.00	.000	−.000	−.000	.650	.241	.087	−.137	.595	.111	−.079	−.028
	2		1.00	.001	.000	−.097	.729	−.229	.037	.033	.644	.296	−.049
	3			1.00	−.001	−.056	.238	.545	−.125	.204	−.081	.274	.182
	4				1.00	.075	.185	.121	.544	.149	.027	.422	−.136
Judge 2	1					1.00	.001	.000	.001	.547	.220	.366	−.117
	2						1.00	−.001	−.002	−.034	.518	−.040	.050
	3							1.00	.000	−.243	.153	.243	.207
	4								1.00	−.025	−.099	.368	.041
Judge 3	1									1.00	−.000	−.000	.000
	2										1.00	.000	−.000
	3											1.00	.000
	4												1.00

Note. All correlations are based on $n = 193$. The validity diagonals are circled.

Judge 3 is involved in these two comparisons suggests that this judge has a somewhat different factorial structure than the other two judges.

This analysis indicates that these three judges generally agree on the clustering of variables, although Judge 3 is somewhat different from the other two. However, the convergent validation results generally indicate a fairly high degree of consensus.

In order to examine the weighting process, the combining function was assumed to be additive (Dawes & Corrigan, 1974). The weights given to the cues were determined by regressing the global judgment on the nine cues, for each judge. Although interpretation of weights in regression is difficult (Darlington, 1968), the intercorrelations of the cues for each judge were generally small.[7] Therefore, problems of multicolinearity were not large in interpreting these weights. The results of this analysis are shown in Table 5.

It can be seen that the ability to model the judge varies among these three judges. The multiple correlation for Judge 2 is low, as compared to the other two judges. What is most important is the fact that the weights for the cues vary greatly across the three judges. Consider Experts 1 and 2. As we have previously seen, these two experts have very similar factor structures, yet, they weight the cues quite differently in coming to their global judgment. As a matter of fact, the cue that is weighted most heavily by Judge 1, Sternberg-Reed cells, is not weighted significantly by Judge 2. Consider the weighting policies of Judges 1 and 3. They do agree that Sternberg-Reed cells should be weighted most highly yet, Judge 3 weights three other cues as being important (benign histiocytes, malignant histiocytes, and plasma cells), while Judge 1 does not. Clearly then, in addition to there being low agreement in fact, there is also low agreement in principle. From our previous results we can say that disagreement in fact is mostly accounted for by disagreement in principle, since the judges seem to code the data in very similar ways.

DISCUSSION

The implications of the results on information weighting for defining expertise are problematic. If one requires that experts have similar weighting policies, the problem of adequately defining *similarity* is immediately raised. For example, Judges 1 and 3 do agree on the importance of Sternberg-Reed cells,

---

[7] The average intercorrelations of the cues (absolute values) are .20, .17, and .22 for Judges 1, 2, and 3, respectively.

TABLE 5

BETA WEIGHTS FOR CUES IN PREDICTING GLOBAL JUDGMENT

Judges	Cues									
	BH	MH	Eos.	Nec.	PC	Neut.	Lymph.	S-R D.	S-R C.	Multiple *R*
1	.00	.12	.11	−.13*	.08	.08	−.16*	.05	.22***	.46***
2	.04	.07	−.13*	.08	−.03	.00	−.05	.09	.09	.27*
3	−.14**	.22***	.07	−.11	.19***	.02	−.07	−.11	.29***	.49***

Note. Results are based on regressions with $n = 193$. BH = benign histiocytes; MH = malignant histiocytes; Eos. = eosinophils; Nec. = necrosis; PC = plasma cells; Neut. = neutrophils; Lymph. = lymphocytes; S-R D. = Sternberg-Reed distribution; S-R C. = Sternberg-Reed cells.
* $p < .10$.
** $p < .05$.
*** $p < .01$.

yet they do not agree on the weighting of the other cues. On the other hand, Judge *2* is clearly different from either Judges 1 or 3. It should be mentioned that Judge *2* was a resident of Judge 1 and was learning this task to some extent. His low correlation suggests another possible criterion for expertise—experts should have at least one significant cue weight. The reason for this is that at least some information in the cues should be used in forming the global judgment.

Although requiring experts to have similar weighting policies would be a very stringent requirement, it would seem that the combining of information lies at the core of expertise. Learning the important relationships between cues and criteria should be one of the most important kinds of learning that is involved in mastering a particular area. Therefore, it is tentatively suggested that experts should exhibit similar weighting policies.

It was stated at the outset that the conditions discussed in the present study were necessary but not sufficient. This clearly implies that other factors are indicative of expertise. Consider our example of businessmen making a forecast of sales for a new product. Whereas they may agree on what some of the relevant attributes of the product are, the real expert may discern/form cues that no one has ever seen before. Similarly, the expert may be the one who can discern and use contingent relationships between cues, that is, complex interactions among cues. These aspects of cue formation and utilization may be an important aspect of expertise.

A second consideration not discussed in this study is how alternatives are first formed in the decision/judgment process. For example, although there may be several ways to market a new product, the expert may be able to elucidate alternatives that have not been thought of before. Again, this is a creative process that we know little about.

A third consideration, somewhat related to the first two, is how the task is *structured*. By this is meant how the problem is defined and the manner in which a solution will be attempted. For example, the process of information search will greatly structure the task. Kleinmuntz (1968) suggests that the expert searches his environment for those cues that yield the greatest amount of information, and this leads to a shortening of the "decision tree" used to reach a judgment. These types of problems have not been discussed in this study since the experiment served to structure the task. However, in nonexperimental settings, this structuring will undoubtedly be of great importance.

A major criticism of the conditions for expertise stated in the present context may be that too much stress has been put on agreement. This agreement has been with respect to clustering, measuring, and weighting cues. As is well known, the history of science is replete with oddballs who did not agree with anyone, yet, were proved to be correct by subsequent events. This criticism has value, yet it conceals the fact that eventually some relevant criterion other than agreement became available (possibly due to new instrumentation or unusual events). However, from a practical point of view, we cannot afford to wait long periods of time to see who may be correct. Actions and decisions have to be taken within a limited time perspective. Therefore, although agreement is by no means the only criterion, it is one that seems relevant to the discussion of expertise.

Where a relevant, or partially relevant criterion can be known within a reasonable time period, expertise can be determined by one's track record. If it is found that two experts do equally well in terms of matching a criterion (and this is significantly higher than some base rate value), yet they do not meet the requirements discussed in the present study (especially that of similar weighting), What are we to say? We can only say that in a highly probabilistic world, there may be many routes to the same goal (cf. Hammond, 1955) and that there may be more than one way to perform the cognitive tasks involved in judgment.

REFERENCES

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin,* 1959, **56,** 81–105.

Chapman, L. J., & Chapman, J. P. Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology,* 1969, **74,** 271–280.

Coppleson, L. W., Factor, R. M., Strum, S. B., Graff, P. W., & Rappaport, H. Observer disagreement in the classification and histology of Hodgkin's disease. *Journal of the National Cancer Institute,* 1970, **45,** 731–740.

Cronbach, L. J., & Meehl, P. E. Construct validity in psycholcgical tests. *Psychological Bulletin,* 1955, **52,** 281–302.

Darlington, R. B. Multiple regression in psychological research and practice. *Psychological Bulletin,* 1968, **69,** 161–182.

Dawes, R. M., & Corrigan, B. Linear models in decision making. *Psychological Bulletin,* 1974, **81,** 95–106.

Einhorn, H. J. Expert measurement and mechanical combination. *Organizational Behavior and Human Performance,* 1972, **7,** 86–106.

Ghiselli, E. E., & Brown, C. W. *Personnel and industrial psychology.* New York: McGraw-Hill, 1955.

Green, D. M., & Swets, J. A. *Signal detection theory and psychophysics.* New York: Wiley, 1966.

Guilford, J. P. *Psychometric methods.* New York: McGraw-Hill, 1954.

Hammond, K. R. Probabilistic functioning and the clinical method. *Psychological Review,* 1955, **62,** 255–262.

Kleinmuntz, B. The processing of clinical information by man and machine. In B. Kleinmuntz (Ed.), *Formal representation of human judgment.* New York: Wiley, 1968.

Naylor, J. C., & Wherry, R. J. The use of simulated stimuli and the "JAN" technique to capture and cluster the policies of raters. *Educational and Psychological Measurement,* 1965, **25,** 969–986.

Nunnally, J. C. *Psychometric theory.* New York: McGraw-Hill, 1967.

Sawyer, J. Measurement *and* prediction: Clinical *and* Statistical. *Psychological Bulletin,* 1966, **66,** 178–200.

Slovic, P., & Lichtenstein, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance,* 1971, **6,** 649–744.

Stanley, J. C. Analysis of unreplicated three-way classifications with applications to rater bias and trait independence. *Psychometrika,* 1961, **26,** 205–219.

Thorndike, R. L. *Personnel selection.* New York: Wiley, 1949.

Thurstone, L. L. *Multiple factor analysis.* Chicago: University of Chicago Press, 1947.

Winer, B. J. *Statistical principles in experimental design.* New York: McGraw-Hill, 1962.

(Received November 12, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 5, 572–577

PREDICTING THE EVALUATION OF REAL STIMULUS OBJECTS FROM ABSTRACT EVALUATION OF THEIR ATTRIBUTES:

THE CASE OF TROUT STREAMS

JORDAN J. LOUVIERE[1]

Department of Geography, Florida State University

Ten employees of Iowa trout hatcheries served as subjects in a study of human judgment behavior. In Phase 1, subjects judged hypothetical trout streams varying in quantity of trout and driving time to stream (5 × 5 factorial design). In Phase 2 subjects judged real Iowa trout streams, and estimated the quantity of trout and the driving time for each stream. These estimates were used in multiple regression equations, derived from the data in Phase 1, to predict the ratings of the real streams in Phase 2. Correlations between predicted and observed judgments ranged from 0.37 (individual subject level) to 0.85 (group level). Results were consistent with empirical work in information processing and nonexperimental work in geography and other spatially oriented sciences.

Individuals daily make judgments which vary in their importance to the functioning of the person. Some judgments may be trivial —what to wear, what to eat—while others may have substantial impact—which car to buy, what courses to take in a given semester, which of several destinations to choose. This study attempts to apply experimental and analytical methodology, used in the study of human judgment behavior, to a practical prediction problem.

The study of human judgment behavior has received increasing attention (see e.g., Slovic & Lichtenstein, 1971). Such behavior involves the operations of weighing and valuation of alternatives. At any point in time, an individual's judgment may be thought of as comprised of two components: (*a*) an evaluation function, that is, the model which the individual employs to weigh and value stimulus entities, and (*b*) the variance in the attributes of these entities over time and space. Much previous work has concentrated on the idiosyncratic evaluation function of judges at one point in time (see Slovic & Lichtenstein, 1971). Within this research paradigm, considerable attention has been devoted to the so-called "expert" judge (Dawes, 1971; Hoffman, Slovic, & Rorer, 1968; Slovic, 1969) whose evaluation function may be assumed to be stable if the alternatives remain stable.

Static models may be very useful in applied work for short-range prediction in which neither the *evaluation* function of the individual nor the *attributes* of the opportunities are expected to change substantially. Hence, an experiment to test the utility of static models for field prediction should focus upon the judgment of entities in a context in which both the individual evaluation functions and the attributes of the entities to be judged can be assumed to remain fairly constant over time. The purpose of this study, therefore, is twofold: (*a*) to describe mathematically the judgment policy (evaluation function) of a sample of expert judges (whose evaluation functions are assumed stable) in a laboratory context in which the attributes of corresponding real world entities are assumed to have remained stable over time, and (*b*) to employ these mathematical policy descriptions to predict a second set of data consisting of the judges' evaluations of corresponding real world stimulus entities.

The real world situation was a problem in recreational planning in which experts are called upon to recommend trout streams for fishermen after evaluation of many relevant sources of information.

---

[1] Requests for reprints should be sent to Jordan J. Louviere, Department of Geography, Florida State University, Tallahassee, Florida 32306.

This study, therefore, contributes to both the subarea of human information processing known as expert decision making and to the application of information-processing methodology to the prediction of behavior in the field.

METHOD

Subjects

The subjects were 10 employees of the three trout hatcheries operated by the State of Iowa. Selection was based upon willingness to participate coincident with being a trout fisherman. Although experiences varied, subjects had a minimum of five years of experience in both work and trout fishing.

Called upon daily by the general public to recommend streams under their management for fishing trips, subjects are required to put together information, experience, etc. about the set of possible destinations (streams) and then to give some measure of their recommendation. These subjects maintain, administer, and/or work at the hatcheries; they constantly monitor and manipulate a number of variables important to stream quality; and they conduct periodic censuses of the trout population. Decisions regarding stocking, stream improvement, etc. are based on this information. The subjects also fish these streams, and thus know them recreationally and professionally.

Task

The judgment task is the recommendation of trout streams as alternative destinations for one-day fishing trips. The subjects are hatchery personnel who are called upon daily by the public to make such judgments. They are experts at making these recommendations and are also expert trout fishermen. Hence, both their evaluation functions and the stimuli to be judged (trout streams) may be considered to be stable. This task, therefore, is assumed to hold both components of the evaluation process constant.

Experimental Procedure

The experimental paradigm parallels that of functional measurement (Anderson, 1970) because it relies upon a factorial design, a numerical response measure, and a within-subjects format. The analytical method, because of its reliance upon multiple regression, is more in accord with that of judgment policy capturing (Hammond, Hursch, & Todd, 1964; Hoffman, 1960, 1968; Stewart & Gelberd, 1972).

The experimental design derives logically from an assumption that judgments about entities (in this instance trout streams) can be described as if the subjects weighed and valued each of a set of relevant attributes (stimuli) of the entities in turn. Thus, one must first select a set of relevant attributes and then present them in combinations to be judged.

Phase 1

A list of stimuli was derived from conversations with the hatchery personnel and other trout fishermen; trout fishing literature was also consulted. Out of a large number of potentially relevant stimuli,[2] travel time and number of trout were selected and varied in a 5 × 5 factorial design (travel time: 15, 60, 105, 150, 195 minutes; number of trout: 2, 14, 26, 38, 50 trout).[3]

The response scale was a 150-millimeter continuous line scale labeled "Recommendation to Fish." Both end anchors and filler items were employed to provide a range of values within which the subjects might couch their judgments and to avoid possible nonlinearities in scale usage. Verbal end labels were "Strong Recommendation Not to Fish" and "Strong Recommendation to Fish," respectively. Subjects responded by making a slash mark at that point on the line which best described their recommendation of the stream described by the informational items. Slash marks were measured to the nearest 5 millimeters by ruler, yielding a 30-category scale.

Each stimulus combination was centered on a separate one third of a sheet of 8½ by 11 typewriter paper, with the response scale centered immediately below. Presentation order of the stimulus items was alternated and the combinations (25) were arrayed in a different random fashion in each of four replications; all replications were placed in a packet. The subjects were instructed to proceed through the packet as carefully as possible, and insofar as possible, not to refer any of their responses to previous items.

Subjects were instructed to imagine that they were in a new state in which they had never fished and were acquainted with a fellow trout fisherman at work. This fisherman offered to supply them with information about a number of *quality* streams in the new state that they could fish in the coming season. The friend asked the subjects to indicate how strongly he would recommend fishing or not fishing each stream during the forthcoming fishing season.

Quality streams were defined as those which had minimally acceptable levels of competing fishermen, size of fish, water temperature, habitat, and environmental aesthetics. To insure that the subjects

---

[2] Three variables initially appeared relevant: number of trout, travel time, and length of fishable stream. Pilot experiments revealed that subjects discounted length of stream, and subsequent discussions indicated that this variable, along with most others, could be held constant.

[3] Values of these variables were chosen from representative data supplied by the Iowa Conservation Commission: 2 trout per quarter of a mile was the lowest count, while 48 was the highest in the spring-stocking surveys. Stream values for real streams range in between these figures, and there is considerable variance. The travel-time values were expressed in hours and minutes: for example, one hour and 45 minutes.

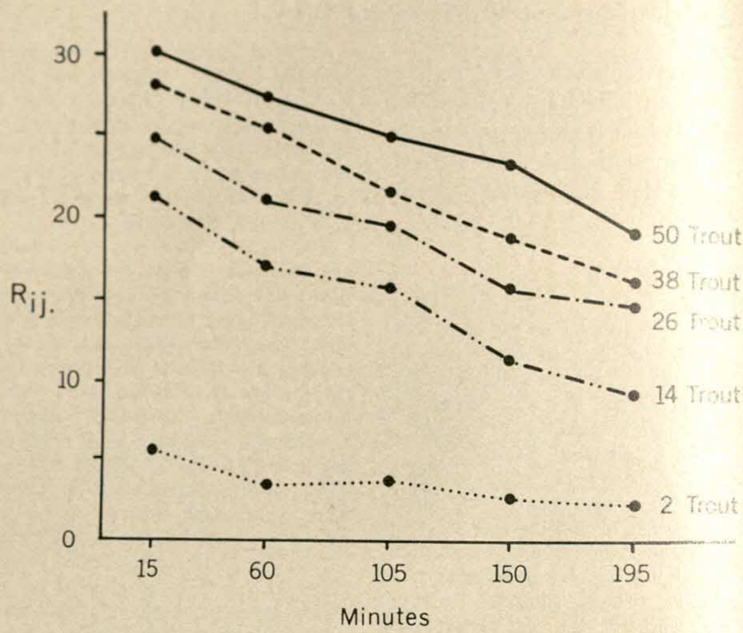


FIGURE 1. Plot of the average rating ($R_{ij.}$) for each stream as a function of number of trout and travel time.

realized that they had to assume these factors were present in each stream, a list was drawn up of the things to assume about each stream. It was left with the subjects so that they could refer to it if need be during the course of the task. A session which averaged 45 to 60 minutes was held at each hatchery. A total of 116 judgments was required [(25 Combinations × 4 Replications) + (4 Fillers × 4 Replications)]. The four fillers represented items that were more extreme than those under experimental investigation. The first replication was treated as practice and was not analyzed.

Phase 2

In the second phase of the study, subjects supplied information about real Iowa trout streams. The purpose was to generate data to be predicted by the equations derived from Phase 1. The subjects first listed up to five of their favorite trout streams. The subjects then rated each favorite stream as they had in Phase 1. Following a stream rating, the subjects supplied a written reason for their rating. Finally, the subjects estimated the average number of trout per quarter of a mile and the average driving time from their residence to each real stream. The subjects were then requested to supply the same information for 10 streams drawn at random from the Iowa trout program.[4]

Analytical Procedure

Analysis of variance (ANOVA) was applied to the response data. Because these data are assumed to have been collected on an interval scale (Anderson, 1972) and because they were observed as the outcome of a repeated measures factorial experiment, hypotheses about the main and interaction effects were tested. The ANOVA served as a test for possible equation forms which would fit the data.

In addition, graphs of the marginal response means plotted as a function of the experimental stimulus values served to compliment the ANOVA results. For the two-factor design in this study, each graph corresponds to a single term in the ANOVA. They supplied visual information suggesting characteristic graphical forms of algebraic expressions (Anderson, 1974; Shanteau & Anderson, 1972), and yielded insight into interaction effects. Both ANOVA and graphical data plots are common analytical procedures of functional measurement (Anderson, 1972).

Multiple regression procedures were employed to derive predictive equations for each individual subject and the group data. These equations were later employed to test their ability to predict the responses of the subjects in Phase 2. Multiple regression is commonly employed in policy-capturing research and has considerable empirical support (e.g., Dawes, 1971; Goldberg, 1968; Stewart & Gelberd, 1972; Yntema & Torgerson, 1961).

---

[4] Depending upon whether one chooses to include experimentally stocked streams or just those part of the regular stocking schedule, there are between 46–60 trout streams in Iowa.

RESULTS

A graph of the response data from the factorial design (Figure 1) suggests that the top four curves follow a linear integration rule; the bottom (2 trout) curve deviates from this form. The latter effect may indicate simply that subjects failed to distinguish this extreme value (2 trout) from the filler value (1 trout) and/or the end anchor value (0 trout). The interaction effect (Trout × Travel Time) was significant ($F = 2.44$, $df = 16/128$, $p < .01$), but localized in the 2 trout curve; removal of this curve rendered the interaction nonsignificant ($F = .76$, $df = 12/96$, $p > .69$). The single factor graphs are in Figure 2. These plots support a linear relationship between the objective and subjective values.[5]

The individual subject graphs exhibit overall similarity to the group graph (Figure 1). All but one subject appeared to follow a linear trend, however, the data exhibited considerable noise. Although a linear predictive function clearly should fit the group data, linear and log-linear functions were fit to the individual subject data for comparison.

The results (Table 1) suggest that the linear models are good predictors of the mean judgmental responses of the individual subjects and the group, although two subjects (5 and 8) deviate from this trend; that is, the log forms fit better. That is not to say that these functions are theoretical models, because they are not, they are predictive functions. The next logical question becomes, How good, in terms of some criterion, are these functions as predictors of Phase 2 data?

Subjects' estimates of the number of trout and travel time (Phase 2) were employed directly to predict their ratings of the real streams (Phase 2). First, the estimates of trout and time were substituted in the multiple regression equations derived from Phase 1 to obtain predicted values for Phase 2 judgments. Correlations between the predicted and the obtained judgments were .37 for the group equation (Table 1: linear group) and .47 for each individual subject's equation (Table 1: Linear, S 1 to S 10). Thus, both equations were used to predict all of the single subject judgments ($n = 68$). The linear model was employed because the plotted group data (Figure 1) were clearly linear and because 8 out of 10 single subjects were better fit by the linear equations.

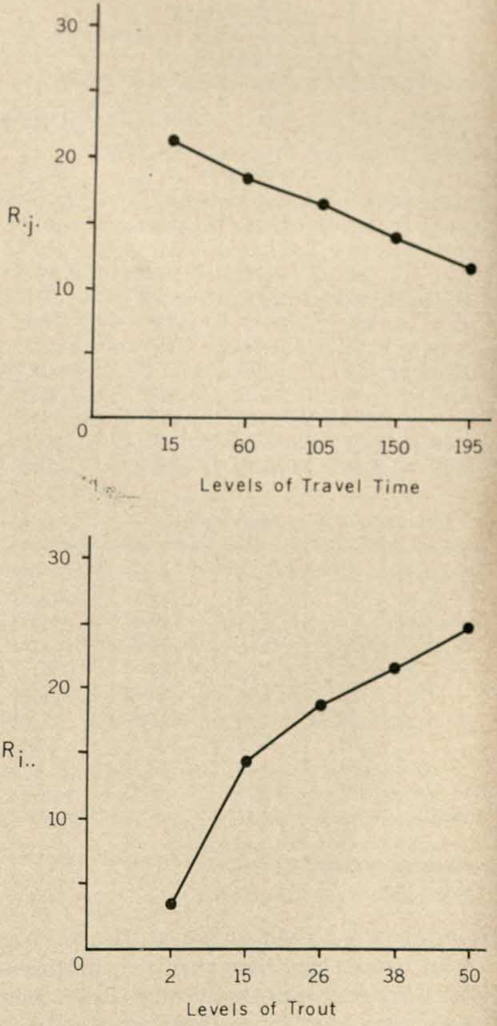


FIGURE 2. Graphs of single variable functions: The average recommendation ($R_{.j.}$; $R_{i..}$) for each value of the number of trout or travel time, holding values of the remaining stimulus constant.

Second, the subject's average estimates of trout and time (Phase 2) were employed to predict the average rating given to each real stream (Phase 2) by the subjects. Subject's average estimates of trout and time were substituted in the group level equations derived

---

[5] A polynomial trend analysis also supports this finding. This, of course, is a necessary assumption for the linear regression model, although Slovic and Lichtenstein (1971) report that the linear model will do well, so long as each independent variable is monotonically related to the dependent variable.

TABLE 1

LINEAR AND LOG-LINEAR FUNCTIONS FIT TO DATA

Subject	R^2	a	B_1 (Trout)	B_2 (Time)
		Linear model		
Group	.97	16.41	.26**	−.04**
S 1	.81	25.13	.21**	−.10**
S 2	.80	18.78	.26**	−.06**
S 3	.90	23.05	.21**	−.11**
S 4	.84	11.35	.45**	−.08**
S 5	.78	9.28	.30**	−.07**
S 6	.79	13.03	.31**	−.04**
S 7	.90	10.36	.40**	−.02**
S 8	.74	6.39	.23**	−.06**
S 9	.58	18.46	.25**	−.02**
S 10	.58	28.36	.04*	−.01**
		Log data model		
Group	.87	1.93	.54**	−.21**
S 1	.48	3.14	.43**	−.38**
S 2	.81	2.52	.42**	−.23**
S 3	.79	3.16	.46**	−.45**
S 4	.83	.89	1.00**	−.37**
S 5	.82	1.29	.90**	−.51**
S 6	.73	1.36	.72**	−.23**
S 7	.89	1.25	.64**	−.10**
S 8	.89	.23	.40**	.30**
S 9	.47	2.14	.38**	−.07
S 10	.49	3.31	.05*	−.03**

Note. *S* = Subject.
* $p < .05$.
** $p < .01$.

from Phase 1 (Table 1: Linear, Group; Log-linear, Group) and these predicted judgments were correlated with the average stream judgments ($n = 12$). Interestingly, the log-linear equations predicted the average rating data better than the linear equation ($R = .75$ compared to .67). This same result obtained when the subject's average estimates (trout and time) were used to predict the average stream ratings by estimating a regression equation employing the average trout and time estimates as independent variables and the average rating for each of the 12 streams as the criterion ($n = 12$). The linear form correlated more poorly (.47) with the average ratings than did the log-linear form (.86). No explanation for this effect can be given at this time.

DISCUSSION

Judgmental Research

Consistent with previous findings in expert decision making (e.g., Dawes, 1971; Stewart & Gelberd, 1972), simple multiple linear regression models accounted for much of the variance in the laboratory data (Phase 1). The application of these functions to the field data must be viewed cautiously, however, because there are a number of possible sources of error that are not strictly detectable. Among these are the following: (*a*) The theory and models may be wrong. (*b*) Some important or relevant stimuli may have been omitted. (*c*) Subjects may not behave in the real world as they do in the lab.

Regarding the first point, even if the true static model were nonlinear, linear models have considerable empirical support (Slovic & Lichtenstein, 1971; Yntema & Torgerson, 1961) for their predictive ability. Concerning the second point, there may indeed be an infinite number of potentially relevant stimuli important to various behaviors, but a most important scientific question is, "How well will only a few do as predictors?" Available evidence (see, e.g., Slovic & Lichtenstein, 1971) suggests that less than four will usually be adequate. Data collected in Phase 2 support this view. Using only the subjects' estimates of trout and time to predict subjects' ratings of real streams, much of the variance in average stream evaluations could be accounted for. Therefore, while there may be other factors operative, these two were sufficient to predict much of the data. Whether subjects behave in the lab as in the real world is a very important question in applied research. The recent presidential polls suggest that individuals will behave consistently with their stated intentions (at least in the short run). However, more research is needed to provide more definitive answers as to the validity and temporal stability of individuals' intentions.

Much research in geography, urban and regional planning, and spatial economics has been directed towards the question of predicting the allocation of trips by individuals over a number of different spatial alternatives (e.g., shopping trips). These studies (e.g., reviewed in Isard, 1959; Olsson, 1965; Wilson, 1968, 1970) have consistently shown that a function of the following form best describes these trip allocations between one or more individuals and one or more points in space:

$$\log M_{ij} = \log C + b_1 \log A_j - b_2 \log D_{ij} + e_{ij}, \qquad [1]$$

where M_{ij} is the observed number of movements (over some time span) between an individual (or an aggregation of individuals) i and a place j; A_j is some objective measure of the attractiveness of place j as a movement destination; and D_{ij} is a measure of the difficulty of moving from point i to point j; C, b_1, and b_2 are constants to be estimated from the data; and e_{ij} is an error term.

This equation, of course, is of the same general form as that derived from the regression of average estimates of trout and travel time against average real stream ratings. Interestingly, the group level log-linear equation of the same form derived from the factorial design (Phase 1) also fit the average ratings in Phase 2 better than its linear counterpart. No explanation can be given for this result, but it may be some artifact of averaging, as the group equation in Phase 1 (Figure 1) seemed to be. Nonetheless, the results are clearly consistent with a large body of research findings in geography and other spatial disciplines cited previously.

CONCLUSIONS

This study has attempted to apply experimental and analytical methodology in human judgment behavior to a practical prediction problem. The methodology appears to be generalizable to a wide range of social problems dealing with human judgment behavior. In particular, it provided insight into the evaluation of alternative destinations which are differentially located in space—a problem of major interest to geography, urban and regional planning, spatial economics, and related disciplines.

A great many theoretical and technical problems await resolution in this type of research. Nonetheless, it is felt that present methodology can be applied to a number of practical problems; moreover, such research can serve as an additional data base for theoreticians. The results of the practical prediction exercise are believed to be significant because they showed encouraging correspondence with data they should have, in theory at least, been able to replicate.

REFERENCES

Anderson, N. H. Functional measurement and psychophysical judgment. *Psychological Review*, 1970, **77**, 153–170.

Anderson, N. H. Looking for configurality in clinical judgment. *Psychological Bulletin*, 1972, **78**, 93–102.

Anderson, N. H. Algebraic models in perception. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 2). New York: Academic Press, 1974.

Dawes, R. M. A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist*, 1971, **26**, 180–188.

Hammond, K. R., Hursch, C. J., & Todd, F. J. Analyzing the components of clinical inference. *Psychological Review*, 1964, **71**, 438–456.

Hoffman, P. J. The paramorphic representation of clinical judgment. *Psychological Bulletin*, 1960, **57**, 116–131.

Hoffman, P. J. Cue-consistency and configurality in human judgment. In B. Kleinmuntz (Ed.), *Formal representation of human judgment*. New York: Wiley, 1968.

Hoffman, P. J., Slovic, P., & Rorer, L. G. An analysis of variance model for the assessment of configural cue utilization in clinical judgment. *Psychological Bulletin*, 1968, **69**, 338–349.

Isard, W. *Methods of regional analysis*. New York: M.I.T. Press, 1960.

Olsson, G. Distance and human interaction: A bibliography and a review. *Bibliography Series No. 2*, Philadelphia Regional Science Association, 1965.

Shanteau, J. C., & Anderson, N. H. Integration theory applied to judgments of the value of information. *Journal of Experimental Psychology*, 1972, **92**, 266–275.

Slovic, P. Analyzing the expert judge: A descriptive study of a stockbroker's decision processes. *Journal of Applied Psychology*, **53**, 1969, 255–263.

Slovic, P., & Lichtenstein, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, 1971, **6**, 649–744.

Stewart, T. R., & Gelberd, L. Capturing judgment policies: A new approach for citizen participation in planning. In J. E. Rickert, C. F. Davis, V. DiMatteo, & A. Bose (Eds.), *Urban and regional information systems: Information research for an urban society* (Vol. I). Claremont, Calif.: Claremont College Printing Service, 1973.

Wilson, A. G. Models in urban planning: A synoptic review of recent literature. *Urban Studies*, **5**, 1968, 249–276.

Wilson, A. G. *Entropy in urban and regional modeling*. London: Pion, 1970.

Yntema, D. B., & Torgerson, W. S. Man-computer cooperation in decisions requiring common sense. *IRE Transactions of the Professional Group on Human Factors in Electronics*, 1961, HFE-2, 20–26.

(Received September 24, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 5, 578–582

STATISTICAL PROCEDURES FOR EVALUATING THE PRACTICAL UTILITY OF A MODERATOR APPROACH TO PREDICTION[1]

ALAN L. GROSS,[2] JANE FAGGEN-STECKLER, AND KAREN McCARTHY

City University of New York

Three statistical procedures for evaluating the practical utility of a moderator approach were described. Each procedure gives tests of significance for the comparison of three prediction strategies: (*a*) using the moderator variable as a basis for subgrouping subjects and computing separate regression equations; (*b*) using the moderator variable as a predictor variable; or (*c*) excluding the moderator variable from the prediction model. It was concluded that when job performance is measured along a continuum, the average *y* score of the selected group (Procedure 1) should be considered. However, when two levels of job performance are measured, for example, pass–fail grades, then either the number of selection errors (Procedure 2) or the number of selection errors when a selection ratio constraint is present (Procedure 3) should be used.

The purpose of this study is to describe statistical procedures that can aid a practitioner in evaluating the practical utility of a moderator approach to prediction. By the expression "moderator approach to prediction" we mean using some variable (moderator) to subgroup a sample of subjects into $g \geq 2$ groups and then computing separate regression equations within each group.

The prediction problem commonly facing a practitioner such as a personnel officer is that of selecting job applicants. In this context the moderator approach has practical utility only if it enables the practitioner to select applicants who are superior in some sense to those who would have been selected had the moderator approach not been used. More specifically, the moderator approach must be compared with two simpler prediction strategies: (*a*) predicting job performance from a single regression equation consisting of p predictor variables ($x_1, x_2, \cdots, x_p$) excluding the moderator (Strategy 1); (*b*) predicting job performance from a single regression equation where the moderator is simply included as the $p + 1$st predictor variable (Strategy 2). Strategies 1 and 2 clearly require the expenditure of substantially less time and effort on the part of the practitioner than does the moderator approach (Strategy 3). Consequently, if either Strategy 1 or Strategy 2 allows one to select a group of job applicants who are equal or superior in some sense to the group selected by the moderator approach, there is little justification for subgrouping subjects and computing separate regression equations.

Let us now consider exactly how the practitioner can assess the superiority, if any, of the group of applicants selected by the moderator approach as compared to Strategies 1 and 2. Following the guidelines suggested by Abrahams and Alf (1972), three statistical procedures are described which enable a practitioner to evaluate the practical utility of the moderator approach to prediction. In Procedure 1, the three prediction strategies are compared in terms of the average performance of the selected group. In Procedure 2, the three prediction strategies are compared in terms of the number of selection errors. In Procedure 3, the prediction strategies are again compared in terms of the number of selection errors, but with the additional constraint that only a fixed number of applicants can be accepted. For each of the three procedures, tests of significance are given for the comparison of the three prediction strategies.

PROCEDURE 1

Assume the investigator is interested in developing a selection model that will lead to the selection of that group of applicants having the highest average job performance (subject to the selection ratio constraint that only n_s applicants can be selected). The investigator has available data for a group of employees on $p \geq 1$ predictor variables $(x_1, x_2, \cdots, x_p)$, a potential moderator variable (m), and a measure of job performance (y). The predictor and moderator scores for these employees should be obtained when these people are applying for employment and, furthermore, should not be used as a basis for selection. The sample is randomly divided into a validation sample and a cross-validation sample. In the validation sample, three different prediction strategies for predicting y are constructed:

---

[1] This research was supported by a grant from the Research Foundation of the City University of New York, 01685.

[2] Requests for reprints should be sent to Alan L. Gross, Graduate School and University Center, City University of New York, 33 West 42nd Street, New York, New York 10036.

Strategy 1—A single least squares regression equation predicting y from the xs is computed.
Strategy 2—A single least squares regression equation predicting y from the x variables and the potential moderator variable m is computed.
Strategy 3—Using the potential moderator variable, the validation sample is subgrouped in some manner into $g \geq 2$ groups and separate regression equations for predicting y from the xs are computed within each group.

In the cross-validation sample, three separate predicted y scores are computed for each subject by applying the validation sample regression weights computed for Strategies 1, 2, and 3 to the cross-validation data. The cross-validation sample predicted y scores for Strategies 1, 2, and 3, respectively are given as

$$\hat{y}_x = b_0 + b_1x_1 + \cdots + b_px_p, \qquad [1]$$

$$\hat{y}_{xm} = c_0 + c_1x_1 + \cdots + c_px_p + c_{p+1}m, \qquad [2]$$

$$\hat{y}_s = \sum_{j=1}^{g} d_j\hat{y}_j, \qquad [3]$$

where d_j is one if a subject is a member of Group j as defined by the moderator variable, and d_j is zero otherwise, $\hat{y}_j$ is the predicted y score for subjects in Group j.

The predicted y score $(\hat{y}_s)$ in the cross-validation sample for the subgrouping strategy (Strategy 3) can best be understood in terms of an example. Suppose $g = 2$ groups have been formed where Group 1 $(d_1 = 1, d_2 = 0)$ consists of all subjects for whom m is less than some value m^*, and Group 2 $(d_1 = 0, d_2 = 1)$ consists of the remaining subjects. The predicted y score for a cross-validation subject for whom $m < m^*$ is obtained by substituting his/her x scores into the validation sample regression equation for Group 1. In a similar manner the predicted y score for Group 2 members is computed.

When there is a positive correlation between y and predicted y in the cross-validation sample, the optimal selection strategy consists of ordering all cross-validation subjects from high to low on the predicted y and then selecting the n_s subjects having the highest predicted y scores (Cochran, 1951). This strategy is optimal in the sense of leading the practitioner to select the group of n_s subjects having the highest average y score.

Using Strategy 1, subjects are ordered on $\hat{y}_x$, and $\bar{y}_x$ (average y score for the n_s subjects having the highest $\hat{y}_x$ scores) is computed. Under Strategy 2 subjects are ordered on $\hat{y}_{xm}$, and $\bar{y}_{xm}$ (average y score for the n_s subjects having the highest $\hat{y}_{xm}$ scores) is computed. Similarly, using the moderator approach (Strategy 3), subjects are ordered on $\hat{y}_s$ and $\bar{y}_s$ (average y score for the n_s subjects having the highest $\hat{y}_s$ scores) is computed. To evaluate the moderator approach the practitioner then compares $\bar{y}_s$ with both $\bar{y}_{xm}$ and $\bar{y}_x$. To justify subgrouping, it must be shown that $\bar{y}_s$ not only exceeds $\bar{y}_{xm}$ but also exceeds $\bar{y}_x$. In other words, we must demonstrate that the average performance of the selected subjects is highest when m is used to subgroup subjects (Strategy 3) rather than used as a predictor variable (Strategy 2) or simply not used at all (Strategy 1). Since the potential for sampling error is greater for a $p + 1$ variable regression equation than a p variable regression equation, it is possible to find in the cross-validation sample that $\bar{y}_s$ exceeds $\bar{y}_{xm}$ but does not exceed $\bar{y}_x$. In this case it is best to predict y from the x variables alone and disregard the moderator completely. If $\bar{y}_s$ does not exceed $\bar{y}_{xm}$, but $\bar{y}_{xm}$ exceeds $\bar{y}_x$, the moderator has practical value only as a predictor variable.

In comparing the three means, $\bar{y}_x$, $\bar{y}_{xm}$, and $\bar{y}_s$, it is useful to have a method for testing the statistical significance of the observed differences. One major difficulty in deriving a significance test is the fact that the n_s y scores used in computing any one of the means are not independent of one another. This dependence follows from the direct ordering of the predicted y scores and the consequent indirect ordering of the y scores (Watterson, 1959). If the y scores are independent, the means of the selected groups may be compared using a nonparametric technique such as the Wilcoxin test (Hays, 1973). However, because of the dependence, this test is only approximate when used to compare $\bar{y}_s$, $\bar{y}_{xm}$, and $\bar{y}_x$.

PROCEDURE 2

In many problems an investigator is interested in developing a selection procedure that will minimize the probability or proportion of selection errors [$P(\mathrm{SE})$]. A selection error can occur in one of two ways: (*a*) selecting an individual whose performance (y) falls below some minimal satisfactory level (y_{crit}); or (*b*) by not selecting an individual whose performance would have equaled or exceeded y_{crit} had he/she been selected. The proportion or probability of selection errors, $P(\mathrm{SE})$, provides a meaningful index for comparing the subgrouping strategy (Strategy 3) with Strategies 1 and 2.

As in Procedure 1, suppose the three different regression equations (Strategies 1, 2, and 3) are computed in the validation sample and each applied to the cross-validation sample. Let us first consider the proportion of selection errors in the cross-validation sample that are incurred by selecting subjects on the basis of their $\hat{y}_x$ scores. Assuming that y and $\hat{y}_x$ are variables that are normally distributed in the cross-validation sample, one minimizes $P(\mathrm{SE})$ if all subjects are selected whose $\hat{y}_x$ scores exceed the cut score, $k(x)$, where

$$k(x) = \frac{s\hat{y}_x}{s_y}\,\frac{(y_{\mathrm{crit}} - \bar{y})}{r(y,\hat{y}_x)} + \bar{\hat{y}}_x \qquad [4]$$

where $s\hat{y}_x$ is standard deviation of predicted y in the cross-validation sample; s_y is standard deviation of y in the cross-validation sample; $\bar{y}$ is mean y score in the cross-validation sample; $\bar{\hat{y}}_x$ is mean predicted y score in the cross-validation sample; $r(y,\hat{y}_x)$ is correlation between y and predicted y in the cross-validation sample.

The derivation of this formula in a standardized form is given by Alf and Dorfman (1967).

To compute the actual proportion of selection errors, resulting from the use of $\hat{y}_x$, $P(\mathrm{SE}|\hat{y}_x)$, that is, prediction Strategy 1, the investigator counts the number of cross-validation subjects for whom $\hat{y}_x \geq k(x)$ and $y < y_{\mathrm{crit}}$ or $\hat{y}_x < k(x)$ and $y \geq y_{\mathrm{crit}}$, and divides this number by the total number of subjects in the cross-validation sample.

The proportion of selection errors incurred in the cross-validation sample using $\hat{y}_{xm}$ as the predictor, $P(\mathrm{SE}|\hat{y}_{xm})$, is computed in the same manner as $P(\mathrm{SE}|\hat{y}_x)$. However, now subjects are selected only if $\hat{y}_{xm} \geq k(x,m)$ where the cut score, $k(x,m)$, is defined by substituting into Equation 4 the required information on $\hat{y}_{xm}$, [i.e., $s\hat{y}_{xm}$, $\bar{\hat{y}}_{xm}$, and $r(y,\hat{y}_{xm})$].

In the subgrouping strategy (Strategy 3) the proportion of selection errors is minimized by separately minimizing the $P(\mathrm{SE})$ value within each subgroup defined by the moderator. Subjects in Subgroup j are selected if their predicted y score ($\hat{y}_j$) exceeds the cut score, k_j, given by

$$k_j = \frac{s\hat{y}_j}{sy_j}\,\frac{(y_{\mathrm{crit}} - \bar{y}_j)}{r_j} + \bar{\hat{y}}_j \quad j = 1, 2, \cdots g, \qquad [5]$$

where $s\hat{y}_j$ is the standard deviation for the predicted $\hat{y}_j$, in Group j; sy_j is the standard deviation of y in Group j; $\bar{y}_j$ is the mean y score in Group j; $\bar{\hat{y}}_j$ is the mean $\hat{y}_j$ score in Group j; r_j is the correlation between y and $\hat{y}_j$ in Group j. In other words, a different cut score is set for each subgroup. The proportion of selection errors incurred by subgrouping, $P(\mathrm{SE}|\hat{y}_s)$, is computed by counting the number of subjects in Subgroup j ($j = 1, 2, \cdots g$) for whom $\hat{y}_j \geq k_j$ and $y < y_{\mathrm{crit}}$ or $\hat{y}_j < k_j$ and $y \geq y_{\mathrm{crit}}$. If we denote this number as N_j, the proportion of selection errors incurred by the subgrouping strategy is given by

$$P(\mathrm{SE}|\hat{y}_s) = \sum_{j=1}^{g} N_j/n_{\mathrm{CVS}} \qquad [6]$$

where n_{CVS} is the total number of subjects in the cross-validation sample.

The subgrouping strategy (Strategy 3) is of practical value only if $P(\mathrm{SE}|\hat{y}_s) < P(\mathrm{SE}|\hat{y}_{xm})$ and $P(\mathrm{SE}|\hat{y}_s) < P(\mathrm{SE}|\hat{y}_x)$. In other words, it must be demonstrated that the proportion of subjects "misclassified" using the moderator to subgroup subjects is less than the proportion misclassified using either the strategy where the moderator is a predictor variable or the strategy where the moderator is not used at all. If $P(\mathrm{SE}|\hat{y}_{xm}) < P(\mathrm{SE}|\hat{y}_s)$ and $P(\mathrm{SE}|\hat{y}_{xm}) < P(\mathrm{SE}|\hat{y}_x)$, the moderator is of practical value only when used as a predictor variable (Strategy 2). If it is demonstrated that $P(\mathrm{SE}|\hat{y}_x) < P(\mathrm{SE}|\hat{y}_s)$ and that $P(\mathrm{SE}|\hat{y}_x) < P(\mathrm{SE}|\hat{y}_{xm})$, one should predict y only from the x variables (Strategy 1).

Testing for the significant differences between $P(\mathrm{SE}|\hat{y}_x)$, $P(\mathrm{SE}|\hat{y}_{xm})$, and $P(\mathrm{SE}|\hat{y}_s)$ is simply the problem of testing the difference between correlated proportions. A simple chi-square test described by Hays (1973) is applicable to this problem.

PROCEDURE 3

In some selection problems, one cannot apply Procedure 2, since the use of cut scores given by Equations 4 and 5 may lead the investigator to select more applicants than there are available "openings." The investigator desires to minimize the proportion of selection errors, but at the same time can only select a certain number (n_s) or proportion (α) of the applicants. Thus the cut scores given by Procedure 2 must be modified.

Using Strategy 1, the n_s subjects having the highest $\hat{y}_x$ scores are selected, and $P(\mathrm{SE}|\hat{y}_x)$ is computed. For Strategy 2 the n_s subjects having the highest $\hat{y}_{xm}$ scores are selected, and $P(\mathrm{SE}|\hat{y}_{xm})$ is computed.

In the subgrouping strategy (Strategy 3), if y and $\hat{y}_j$ are variables that are normally distributed in subgroup j $(j = 1, 2, \cdots, g)$ $P(\mathrm{SE}|\hat{y}_s)$ is minimized subject to the selection ratio constraint if subjects in Subgroup j are selected only if $\hat{y}_j$ equals or exceeds k_j. The g cut scores, $k_1, k_2, \cdots, k_g$, are computed by iteratively solving the following equations:

$$P_1 = P_2 = \cdots = P_g \qquad [7]$$

$$\alpha = \sum_{j=1}^{g} q_j P(Z \geq k_j^*) \qquad [8]$$

where the probability P_j is given as

$$P_j = P[Z \geq (y^*_{crit(j)} - r_j k_j^*)/(1 - r_j^2)],$$

where Z is a unit normal variable, $y^*_{crit(j)}$ is the value of y_{crit}, standardized in Group j, r_j is the correlation between y and $\hat{y}_j$ in Group j, k_j^* is the value of k_j, standardized in Group j, q_j is the proportional size of the jth subgroup in the cross-validation sample relative to the entire cross-validation sample, α is the proportion of individuals in the cross-validation sample to be selected.

The derivation of Equations 7 and 8 is given by Gross and Su (1973).

To better understand the computation of the cut scores given by Equations 7 and 8, consider an example where there are $g = 2$ subgroups in the cross-validation sample. In subgroup j ($j = 1$ or $j = 2$) the mean y score ($\bar{y}_j$), the mean predicted y score ($\bar{\hat{y}}_j$), the standard deviation of y (sy_j), the standard deviation of predicted y ($s\hat{y}_j$) and the correlation between y and predicted y (r_j) can be computed. If there are n_j subjects in the jth group, q_j is computed by dividing n_j by the total number of subjects in the cross-validation sample. To standardize y_{crit} in group j ($y_{crit}{}^*{}_{(j)}$), subtract $\bar{y}_j$ from y_{crit} and divide by sy_j. The standardized value of k_j (k_j^*) is similarly obtained by subtracting $\bar{\hat{y}}_j$ from k_j and then dividing by $s\hat{y}_j$. To find k_1^* and k_2^* a table of unit normal probabilities is consulted and values are selected that make P_1 equal to P_2 (i.e., satisfy Equation 7) and also yield a selection ratio of α, (i.e., satisfy Equation 8). Once the standardized values for the cut scores (k_1^*, k_2^*) are found, the raw score values (k_1, k_2) are readily computed. As in Procedure 2, the values for $P(\mathrm{SE}|\hat{y}_x)$, $P(\mathrm{SE}|\hat{y}_{xm})$ and $P(\mathrm{SE}|\hat{y}_s)$ are then compared using the chi-square test for correlated proportions.

DISCUSSION

Any one of the three procedures described in the previous section can readily be used by a practitioner in determining the practical utility of the moderator approach. The type of selection problem facing the practitioner usually dictates which of the three procedures is most appropriate. In situations where job performance is measured along some type of continuum, it is meaningful to consider the average y score of the selected group in evaluating the moderator approach (Procedure 1). When only two levels of job performance are measured ($y \geq y_{crit}$ or $y < y_{crit}$) such as pass–fail grades, then either Procedure 2 or 3 could be used.

REFERENCES

Abrahams, N. M., & Alf, E. F. Pratfalls in moderator research. *Journal of Applied Psychology*, 1972, **56**, 245–251.

Alf, E. F., & Dorfman, D. D. The classification of individuals in two criterion groups on the basis of a discontinuous payoff function. *Psychometrika*, 1967, **32**, 115–123.

Cochran, W. G. Improvement by means of selection. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 1951, 449–470.

Gross, A. L., & Su, W. H. A decision theory approach to test bias. Unpublished manuscript, City University of New York, Department of Education, 1973.

Hays, W. L. *Statistics for the social sciences*. New York: Holt, Rinehart & Winston, 1973.

Watterson, G. A. Linear estimation in censored samples from multivariate normal populations. *Annals of Mathematical Statistics*, 1959, 30, 814–824.

(Received August 1, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 5, 583–595

UTILIZATION OF MANPOWER:

DEVELOPMENT AND EVALUATION OF AN ASSESSMENT-CLASSIFICATION MODEL FOR MATCHING INDIVIDUALS WITH JOBS[1]

LYLE F. SCHOENFELDT[2]

University of Georgia

A model for the assessment of individual characteristics, the identification of psychological requirements of jobs, and the classification of applicants to job opportunities was developed. The validity of the model was investigated with a large sample of students ($N = 1,934$) working toward a college degree. Subgroups, formed on the basis of previous behavioral data collected during the freshman year, differed with respect to criteria (arts-sciences vs. applied studies, grade point average, etc.) measurements taken four years later. More importantly, the subgroups differed with respect to the curricular paths walked during college. The results indicated that it was possible to differentiate people in meaningful ways, to identify "job families," and to match people with jobs.

A recent trend has been the emergence of a feeling shared by a growing chorus of behavioral scientists and the public that it should be possible to do a better job of predicting human behavior. A complicating factor for industrial/organizational psychologists is the observation by Meyer (1972) that the world of work is undergoing profound and rapid change. The dual dilemma of (*a*) the ceiling on our ability to predict behavior and (*b*) the rapidity of change is receiving increasing attention in varied publications (Owens, 1970) and contexts, for example, clinical psychology (Cole & Bruner, 1971), experimental psychology (Vale & Vale, 1969), the educational setting (Novick & Jackson, 1970), and, of course, the industrial setting (Barrett, 1972; Campbell, Dunnette, Lawler, & Weick, 1970).

Interest in this dilemma represents a growing dissatisfaction with the state of the art; a dissatisfaction which culminated in the formation of a model directed toward (*a*) the *assessment* of individual characteristics, (*b*) the identification of the psychological requirements of jobs, and (*c*) the *classification* of manpower to available opportunities. The purpose of the present research was to review the evidence relative to the assessment–classification model and to evaluate the potential gains offered by this approach in the context of manpower utilization.

As far back as 1948, Toops (1948) discussed the value of analyzing relationships in terms of homogeneous groups. Knowledge of an individual's group membership enabled the psychologist to predict future behavior with increased facility (Toops, 1959). Toops's line of reasoning has become the intellectual antecedent for a class of systems prediction models. The approach of *systems models* in contrast to that of statistical models (exemplified by the work of Ghiselli, 1956, 1960; Lee, 1961) is typically more complex and nebulous, or at least subject to varying interpretation by differing researchers.

One line of descent from the Toops (1948) conception has been the *person–process–product* models by Dunnette (1963), Campbell et al. (1970), and Cirino-Gerena (1972) emphasizing the individual's configuration of differences and the impact this has on job effectiveness. The models put forth by each of these researchers view the product as a result of the configuration, that is, a person inter-

---

[1] This research was supported in part by the National Institutes of Child Health and Human Development under Research Grant 5-R01-HD04135-03.

[2] The author gratefully acknowledges William A. Owens and Hubert S. Feild for their comments and suggestions on this research.

Requests for reprints should be sent to Lyle F. Schoenfeldt, Department of Psychology, University of Georgia, Athens, Georgia 30602.

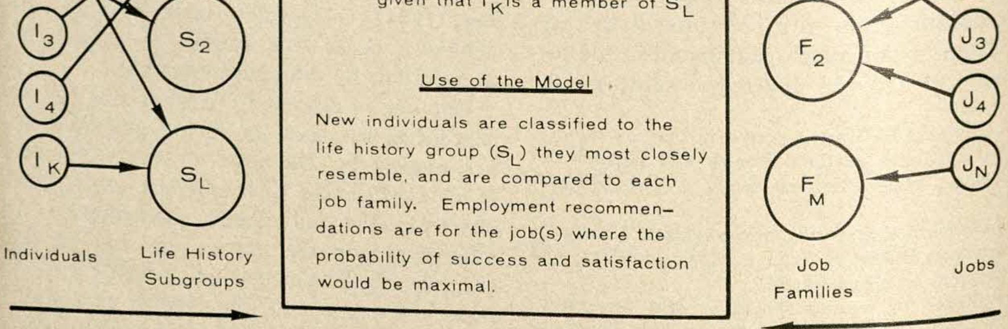
acting with various task demands or organizational forces.

A second line of descent from Toops was offered by Owens (1968). His *developmental–integrative model* was based on the premise that past behavior is the best predictor of future behavior. As with the other schemes, the point of entry was a configuration of individual differences, except in his model the quantification of these differences was heavily laced with the measurement of antecedent or life-experience characteristics. Prediction was to be accomplished by forming subgroups with respect to previous experiences and relating the subgroups to criteria of interest. Subgroups would be expected to display differential profiles across the criterion behaviors. It must be added, however, that the individual was not lost in this conception. Owens (1971) had fully detailed the implications of his approach for individual assessment.

The *assessment–classification model,* shown in Figure 1, follows from the other systems approaches. It incorporates the evaluation of person–process–product models with the subgroup conception of Owens (1968). It involves the assessment of individuals, measurement of jobs, and the prediction of job success.

Individual assessment or the inventory of the psychological capabilities the individual brings to the job market has two aspects. The first involves using standard predictors found to be valid for the jobs in question, the individual differences variables of the Campbell et al. (1970) model. The second aspect involves implementation of the approach described by Owens (1968) involving the formation of subgroups with respect to the major dimensions of antecedent behavior and relating the subgroups to relevant criteria. This would involve administering a biographical questionnaire to assess the antecedent behaviors. The individuals would then be classified on the basis of their responses to life history subgroups or subsets homogeneous with respect to important dimensions of life behavior.

After lengthy neglect, job analysis is seen as a way of understanding the nature of work behavior (Prien & Ronan, 1971). Several instruments (Cunningham, Tuttle, Floyd, & Bates, 1970; McCormick, Jeanneret, & Mecham, 1969) and operational procedures (Fleishman, 1967) have been developed and found to be of use for measuring jobs in terms of the psychological demands required for successful performance. Thus, in the same way individuals are placed in subgroups homogeneous with respect to past behavior, jobs can be classified into families homogeneous with respect to their psychological requisites (Taylor, 1970). This would involve the



Establishment of the Model

Discriminant analysis to determine the probability of success and satisfaction in F_M given that I_K is a member of S_L

Use of the Model

New individuals are classified to the life history group (S_L) they most closely resemble, and are compared to each job family. Employment recommendations are for the job(s) where the probability of success and satisfaction would be maximal.

FIGURE 1. The assessment–classification model of manpower utilization.

analysis of the job by a technique such as the Position Analysis Questionnaire (McCormick et al., 1969) and the classification of jobs based on the responses to the analysis into job families homogeneous with respect to important activities (and hence job requirements). The assessment–classification model is then developed by the use of a maximization procedure (such as discriminant analysis) to determine the probability of success and satisfaction in a particular job family given that the individual is a member of a particular life history subgroup. The goal of the model is the prediction of job success and satisfaction for a new individual whose subgroup is known and who would be performing a job belonging to one of the job families.

Some evidence that the assessment–classification approach to prediction is feasible came from Novick and Jackson (1970). These authors concluded that the logical approach is a Bayesian model. Thus, although the usual predictive procedures (linear regression) have, at least, intermediate effectiveness, the overall effectiveness can be increased if recognizable subgroups, deviant from the average in predicted success, can be removed. In contrast to statistical models for improved prediction, which have resulted in a mixed picture of successes and failures (Zedeck, 1971), the assessment–classification model has never been given a full and complete evaluation.

This article describes a research effort to investigate the usefulness of the assessment–classification model. The evidence, hopefully, will point to a successful generation of new predictive procedures and increased understanding of job related behavior. Specifically, this article develops an assessment–classification model in a college situation and explores the stability and validity of that model.

Most of the systems models have been concerned with prediction of work performance, usually work at the managerial level. However, it seems appropriate to use college students in a study of the relationship of human differences to job effectiveness. College students have much in common with job incumbents. In each case (college or job) the individual brings with him (or her) a configuration of differences. The course (as well as the subelements within the course) is equivalent to task demands—demands which result in behavior that is inevitably evaluated as to its effectiveness. The educational processes occur within the context of an organizational climate which offers constraints of various types. The college reward structure has many levels, several facets at each level, and if nothing else, is well imbued on the individuals. In the college setting, as in an organization, there is feedback, both immediate and delayed, as to the effectiveness of behavior. Finally, for both tasks and courses, training and development experiences are optional and may or may not occur. In addition to these similarities, the advantage of the educational setting of permitting a broader range of research studies than would otherwise be possible in the industrial context more than offsets the assumed loss in generalizability of results to the world of work.

METHOD

Subjects

The subjects were 1,037 males and 897 females, an 80% sample of the 1968 freshman class at the University of Georgia. The study was longitudinal extending over a four-year period.

Biographical Data

The biographical data were collected during the first quarter in residence. A detailed discussion of the methodology regarding the development of the biographical inventory is available elsewhere (Schoenfeldt, 1970). The goal was to identify subsets of subjects with comparable patterns of prior experience. A 389-item biographical inventory was constructed such that each item was multiple choice, with the responses arranged to form a continuum. Scales were formed from the items by separate principal-components analyses and varimax rotations for the sex groups, which resulted in 13 (Table 1) and 15 (Table 2) interpretable factors, respectively.

Subgroups

Each subject was profiled on the life-experience dimensions and each profile was then transformed into a matrix of intersubject profile similarities. Profiles were grouped into "families" using the procedure developed by Ward and Hook (1963). One deficiency in the grouping procedure has been that once assigned to a group, the individual remains in that group, thus the assignment of individuals to subsets is usually less than optimal at the conclusion of the grouping (Ward, 1963). A two-part

TABLE 1

DESCRIPTION OF MALE BIOGRAPHICAL FACTORS

Factor	R[a]	ρ_c[b]
1. Warmth of Parental Relationship Had a close, warm relationship with both parents; parents were very likely to give affection, praise and attention; wanted to imitate and be like the father.	.96	.94
2. Academic Achievement High standing in high school class; often on the semester honor roll; very competitive and successful in academic situations.	.96	.95
3. Social Introversion Compared to others in high school: fewer casual friends, fewer dates, less popular, more self-conscious and ineffective in meeting demands of social situations.	.95	.94
4. Athletic Interest Very active in athletic activities; often engaged in team sports; enjoyed physical education courses; rated past performance in physical activities very high.	.97	.96
5. Intellectualism Regularly read literary, business, or scientific magazines; watched educational and cultural TV shows.	.93	.91
6. Verbal Aggressiveness Enjoyed discussion courses, and tended to try to make others see their point of view; questioned teachers on subject matter a lot; were often regarded as radical or unconventional; often wanted to be alone to pursue own thoughts and interests.	.94	.93
7. Socioeconomic Status High parental educational level, average family income, and father's occupational level is high.	.96	.95
8. Parental Control vs. Freedom (bipolar) Parents were more strict, critical, or punitive; anger was more often shown by or at parents; were allowed much less freedom or independence.	.94	.92
9. Positive Adjustment Response Bias Rarely wished to become more socially acceptable; less often chose parents or friends as someone to "take things out" on; less often felt downcast, dejected or selfconscious; not typical to daydream.	.94	.93
10. Scientific Interest Enjoyed science and lab courses, and found them relatively easy; worked with scientific apparatus and equipment, often outside of any required school assignment.	.95	.94
11. Positive Academic Attitude Liked school and teachers very much; teachers were more successful in arousing academic interests; enjoyed specific courses more; did more hours of homework.	.94	.92
12. Religious Activity Very active in church, religious or charitable organizations; compared to others of the same age, more often went to church, and had stronger religious belief.	.91	.88
13. Sibling Friction More often argued or fought with siblings; had more younger brothers and sisters, close to their own age; more friction and feelings of competition.	.96	.95

Note. Biographical factors are listed in descending order of variance explained.

[a] Multiple correlation using 118 biographical items (short form) to predict each factor score.

[b] The unbiased estimate of the cross-validated population correlation, which is the multiple correlation obtained by applying the sample weights to the entire population (Claudy, 1970).

procedure was developed to evaluate the fit of each individual to the assigned group and remove him from the structure if he failed to fit any group well (isolate), or fit more than one group about equally (overlap) (Schoenfeldt, 1970).

Using this procedure, 23 male and 15 female subgroups were formed. Approximately 75% of the total sample was allocated to one of the groups with the remaining subjects either not fitting any of the groups or, alternately, matching two or more groups.

As might be expected, differences between the groups on the biographical factors were highly significant, thus allowing the characterization of each group in terms of the dimensions that distinguished it from all or most of the remaining groups. In addition, each subject had completed a battery of standard and experimental tests when originally surveyed in 1968. These measures have been described by Owens (1971), and included, among other things, the Scholastic Aptitude Test (SAT), the Strong Vocational Interest Blank (SVIB), the California F

Scale, the Inventory of Social Attitudes (tough-mindedness versus tender-mindedness and conservative versus radicalism), the General Outlook Inventory (integrative complexity, conceptual simplicity, and hierarchical complexity), and the Purdue Values Inventory (cognitive values, economic values, and social–religious conformity). Finally, these subjects had been involved in a number of field studies, such

TABLE 2

DESCRIPTION OF FEMALE BIOGRAPHICAL FACTORS

Factor	R^a	$\rho_c{}^b$
1. Warmth of Maternal Relationship In high school, very close to the mother; mother more often provided emotional support and interest; more often discussed intimate and/or important matters with mother.	.94	.92
2. Social Leadership More often tended to guide or direct others in group activities; participated in school politics; held more leadership positions in high school.	.95	.94
3. Academic Achievement High grades; very high standing in high school class; more competitive in academic situations; more often expected to be successful in academic tasks.	.96	.95
4. Socioeconomic Status High educational and occupational level of father high family income and social class.	.96	.95
5. Parental Control vs. Freedom (bipolar) Parents were more strict, and allowed less freedom or independence, and were more punitive by taking away privileges.	.93	.91
6. Reading Interest Regularly read national news magazines; did much more nonrequired reading; often watched TV news programs and special reports.	.88	.85
7. Athletic Participation Rated past performance in physical activities very high, very active in athletic activities; more often engaged in individual and team sports, enjoyed physical education courses.	.97	.96
8. Scientific–Artistic Interests Enjoyed courses in the sciences, music, or art, more often worked with scientific equipment or apparatus; physical and biological science subjects; active in dramatic, art or music groups.	.92	.90
9. Conformity to Female Role More often suffered "attacks of conscience" when they felt they had done wrong by the standards of society, the church or parents; more often wished to become more socially acceptable, or better prepared as a responsible family member.	.88	.84
10. Maladjustment Often typically felt downcast and dejected, felt so upset that they brooded over the meaning of life; daydreamed to a greater extent; were more sensitive to criticism.	.92	.90
11. Expression of Negative Emotions Very often openly expressed anger with a close friend; very often tried to get even when someone close hurt or upset them; more often wanted to "take things out" on friends.	.87	.84
12. Social Maturity More likely to give help or advice to friends with personal problems; enjoyed complete freedom to work as they pleased in class projects or tasks; were respected by classmates.	.86	.82
13. Popularity with the Opposite Sex More often went on dates; started dating regularly and started going steady at younger age.	.90	.88
14. Positive Academic Attitude Teachers were more successful in arousing academic interests, and allowed much class participation and discussion; liked high school teachers to a greater degree; felt high school education was adequate.	.92	.90
15. Daddy's Girl Were very close to the father, and spent relatively more time with him, father provided emotional support, interest, and gave more attention.	.92	.90

Note. Biographical factors are listed in descending order of variance explained.

[a] Multiple correlation using 118 biographical items (short form) to predict each factor score.

[b] The unbiased estimate of the cross-validated population correlation, which is the multiple correlation obtained by applying the sample weights to the entire population (Claudy, 1970).

as leadership (Boardman, Calhoun, & Schiel, 1972) and over- versus underachievement (Klein, 1973). Tables 3 (males) and 4 (females) summarize the subgroup differences in terms of these several sets of variables and research results. For example, Group 1 of the males is distinguished by virtue of the fact that its members scored extremely low on the dimensions of intellectualism, athletic interest, and verbal aggressiveness. Group 1 males were also distinguished from the remaining 22 groups in that they scored high on the Music Teacher scale of the SVIB. Groups that were high on a particular life-experience dimension generally scored in the same direction on comparable tests from the supplemental battery, although there was not always room on the table to indicate this duplicative information.

Stability

A critical aspect of the logical flow concerns the information that the life-experience dimensions and subgroup structures have been found to be stable in several important respects. In 1970, 1971, and 1972, over 6,000 additional subjects completed 118 of the biographical data items with the highest loadings on the factors. Using the estimated factor scores computed from the 118-item form (see Tables 1 and 2 for the multiple correlations), each new subject was classified to the group he (or she) most resembled by means of using the discriminant-classification procedure described by Cooley and Lohnes (1971).

Stability was assessed by comparing the allocations of subjects to groups for the years subsequent to 1968, as well as by comparing the allocations of these years to the allocations of 1968. With regard to 1970 through 1972, the difference in the percentage of subjects classified to each group exceeded 3% in only one of the 38 (23 male and 15 female) comparisons, the exception being female Group 13 which had 6% of the subjects in 1970–1971 and 10% in 1972. For 1970–1972, 35% (versus 27% in 1968) of the males and 32% (versus 26%) of the females were designated as overlaps or isolates. Other than these differences, only two of the 38 groups formed in 1968 differed by more than 4% in 1970–1972; male Group 3 decreased by 6%, while female Group 10 increased by the same amount. In summary, the subgroup structure established with subjects tested in 1968 was replicated during the three subsequent years. The subgroup structure was wholly applicable to other samples surveyed at the same institution over a four-year period.

Another finding has been that the life-experience dimensions were not specific to sex or institution (Anderson, 1973). This study confirmed that finding. Approximately 5,400 subjects representing three diverse universities completed a large number of biographical items, 248 of which were selected for analysis, including sex and institution which were coded as dummy variables. The item responses were factor analyzed as described previously, except that sex and institution were forced to become the first two factors (Cooley & Lohnes, 1971). There was no meaningful institutional variance. In addition, it was possible to define a factor that included all of the sex variance, and then describe a set of 16 additional factors (similar to those described in Tables 1 and 2) that cut across both sexes.

The successful application of the assessment-classification model hinges not only on the formation of stable, meaningful subgroups on the basis of antecedent behavior but also on the relevance of the subgroups to job criteria. The subgroups must mesh with groups formed on the basis of job activities.

Validating the Model

The validation of subgroup differences was with respect to three types of educational criteria. First, group differences were evaluated on five univariate criteria. Second, course profile subgroups were formed and related to the life-experience groups. The third analysis was concerned with group differences on the components of predicted and grade point average, that is, the SATs and high school grades, to evaluate the generalizability of results over a two-year period. All educational criteria were obtained from transcript records of the 1,934 subjects tested in 1968.

A count was made for each subject of the number of courses (i.e., job activities) completed in each of the following 12 areas: (*a*) fine arts, (*b*) languages, (*c*) humanities, (*d*) biological sciences, (*e*) physical sciences, (*f*) social sciences, (*g*) agriculture, (*h*) business, (*i*) education, (*j*) home economics, (*k*) speech–journalism, and (*l*) physical education–ROTC. In the same manner as was used to form life-experience subgroups, course profile subgroups were derived for the 1,450 subjects (sexes combined) who had completed 90 or more quarter hours over a period of 13 quarters (summer included) following matriculation. The result was 11 course profile subgroups or job families including 69% of the subjects and corresponding to the 12 course areas except that the biological and physical sciences were combined. A twelfth group, fewer hours group, was distinguished by the fact that they averaged 47 fewer credit hours than the remainder of the subjects (although all completed more than 90 hours).

RESULTS

Criterion Differences

The 23 male and 15 female life-experience subgroups were compared on five educational criteria. Chi-squares were computed comparing for each group the number who had (*a*) been on the dean's list one or more times versus the number on at no time; (*b*) been on academic probation and/or dismissal one or more times versus at no time, and (*c*) had one or more course withdrawals versus no withdrawals.[3] The 11 course major profiles were divided into arts and sciences versus applied studies, and a chi-square was used to compare the subgroups on this dichotomous criterion. The final criterion was grade point average, and the groups were compared on this variable by means of a one-way analysis of variance (ANOVA).

The results are presented in the center columns of Tables 3 (males) and 4 (females). The average percentages and chi-squares (or *F*) are provided at the base of each column. Groups deviating substantially from the average value are marked as high or low. Withdrawals for the females was the only criterion on which the groups did not differ significantly.[4] As can be seen from the tables, the principal interrelationships between the life-experience subgroups and the criteria were with respect to dean's list awards, probations, and grade point average. The remaining criteria, withdrawals and arts–sciences, were relatively independent of life-experience subgroup membership.

Assessment–classification Relationships

It was of interest to see how the antecedent behavior subgroups related to the course profile (or job family) subgroups. To answer this question, two biographical data by course profile group matrices were formed, a 23 × 9 for males (languages, education, and home economics were null sets) and a 15 × 10 for females (agriculture and physical education were null sets). The highlights of each row of these matrices are described in the right-most columns of Tables 3 and 4.

As a chi-square assumption could not be met, that is, more than 20% of the expected values were less than five, two alternate procedures were used. The first technique, developed by Kundert (1973), in conjunction with a Fisher-Lancaster categorical scaling program, involved the calculation of a chi-square and, then the generation of a random matrix with the same marginals as in the original data matrix. For each observed matrix, a random data table was constructed and a chi-square computed. Thus, for each analysis an empirical sampling distribution was generated and related to the observed results. One limitation was that the program could only handle a 10 × 10 matrix, necessitating the formation of 10 observed matrices from the original 23 × 9 and three from the 15 × 10 such that each pair of groups were included in at least one of the analyses. For the males, six of the 10 chi-squares were significant at less than the .05 level, and eight at less than the .08 level. All three of the chi-squares calculated from the female data were significant at less than the .01 level.

A second feature involved the assignment of dummy variable codes to the categories, which were then used for calculating a canonical correlation between the biographical data and course profile subgroup categories. These codes also served as input data for computing a discriminant function between the biographical data subgroups (with course profiles as criteria) and between the course profiles (with biographical data groups as criteria). This information was then used for determining which rows or columns were accounting for differentiation between the groups in the data matrix. For the males, science and business contributed most to the linear combination of the course profile groups in the discrimination between the 23 biographical data groups, while biographical data groups 3, 21, and 16 (in that order) contributed most of the differentiation between the course profile groups. The science and fewer hours groups contributed most to the female course profiles in the discrimination between the 15 biographical data groups, while biographical data groups 2, 7, and 12 contributed (equally) most to the differentiation between the course profiles.

---

[3] Among the students with greater than 90 hours, those students with one dean's list, withdrawal, or, to a lesser extent, probation were likely to have several. Using the total dean's lists, for example, rather than the number with one or more exaggerates the strength of the dependent relationship between subgroup membership and the criterion.

[4] Dropouts resulted in restriction of the range of criterion scores, thus diminishing the strength of subgroup-criterion relationships to an unknown extent. Previous results (Schoenfeldt, 1970) showed that dropouts were not evenly distributed across subgroups, but that some groups had less than 10% dropouts after one year, while the rate was as high as 50% for others.

TABLE 3

SUBGROUP DIFFERENCES ON INDIVIDUAL DIFFERENCES DIMENSIONS AND SELECTED CRITERIA—MALES

Group	Subgroup characteristics in terms of biographical factors, test battery, and field studies[a]	1968 *n*	Subgroup differences on criteria[b]					Subgroup characteristics in terms of course profile
			$\% \geq 1$ dean's list	$\% \geq 1$ probation or dismissal	$\% \geq 1$ course with-draws	Grade point average	% arts–science	
1	Low Intellectualism, low Athletic, low Aggression ** high Music	44	26	44	44	2.47	53	10% fine arts (vs. 3%)[c]
2	High Achievement, low Introversion ** high F scale, low Radicalism ++ high Leader	48	44–high	31–low	52	2.75	55	28% science (vs. 16%)[c]
3	High Intellectualism, low Aggression ** low SAT, high Simplicity, low Complexity	73	21	65–high	53	2.10	9–low	49% business (vs. 24%)[c]
4	Low Religion, low Athletic, high Adjustment ** high Librarian ++ high Leader, Overachiever	33	52–high	21–low	30–low	2.76	39	26% journalism (vs. 10%)[c]; 26% science
5	Low Sibling Friction, low Adjustment ** high Armed Forces Officer, low Advertising	31	23	45	71–high	2.33	33	
6	High Religion, low Introversion, high Aggression, low Science ** high Complexity	17	18	18–low	71–high	2.65	50	
7	High Introversion, low Control, low Athletic ** high SAT, low Radicalism ++ high Leader	14	62–high	8–low	38	3.16–high	70–high	50% science
8	High Academic Attitude, high Status ** high People Interest	48	21	44	48	2.51	64–high	28% science
9	High Intellectualism, high Control, low Religion ** high F scale, low SAT, high People Interest	34	24	70–high	67–high	2.19	40	35% business; 22% social science (vs. 11%)[c]
10	Low Religion, high Sibling Friction, high Athletic, low Science ** high F scale, low Complexity	29	11–low	41	52	2.30	44	22% physical education (vs. 10%)[c]
11	High Aggression, high Sibling Friction, low Academic Attitude ** high Complexity, low F scale, low Economic Values	25	39	30–low	57	2.59	38	

TABLE 3 (*Continued*)

Group	Subgroup characteristics in terms of biographical factors, test battery, and field studies[a]	1968 *n*	Subgroup differences on criteria[b]					Subgroup characteristics in terms of course profile
			% ≥ 1 dean's list	% ≥ 1 probation or dismissal	% ≥ 1 course with-draws	Grade point average	% arts–science	
12	Low Religion, high Athletic, low Science ** high Insurance Sales ++ Underachiever	42	36	38	31–low	2.59	21–low	24% physical education; 36% business
13	Low Status, high Adjustment ** low Radicalism	58	17	55	41	2.25	39	18% agriculture (vs. 6%)[c]
14	Low Athletic, high Science, low Sibling Friction ** high Thing and Unstructured Interests	27	42	31–low	35	2.58	62–high	29% social science
15	High Science, high Athletic, low Status	25	32	44	40	2.39	67–high	40% science
16	High Adjustment, high Introversion, high Warmth ** high Thing Interest ++ Underachiever	24	26	43	48	2.48	60	
17	High Achievement, high Status ** high SAT ++ Overachiever	22	77–high	14–low	55	3.31–high	31	
18	Low Adjustment, high Control, high Introversion	22	27	50	45	2.39	36	46% business
19	High Control, high Academic Attitude, high Aggression, low Warmth ** high Thing Interest	14	15	54	62	2.33	100–high	44% fewer hours group (vs. 11%)[c]
20	Low Religion, high Science, low Achievement, high Athletic ++ Underachiever	45	7–low	71–high	59	2.13	55	
21	Low Aggression, high Achievement, high Introversion ** low Complexity, high Thing Interest	25	60–high	36	28–low	2.88–high	67–high	47% science
22	High Control, low Warmth, low Introversion	37	11–low	75–high	53	2.08–low	35	50% business
23	Low Achievement, low Religion, low Sibling Friction ** low F scale, low Complexity	24	8–low	67–high	58	1.96–low	40	29% fewer hours group
Average			28%	47%	49%	2.43	45%	
χ^2 or *F*			97.57***	89.94***	38.04*	7.25***	49.64***	

---

[a] Differences on the biographical inventory are listed first. Differences on the test battery are preceded by two asterisks (**) while field study differences are preceded by two plus signs (++).

[b] Groups deviating substantially from the average value in an upward direction are marked high, those deviating in a downward direction are marked low.

[c] Percentage of males in all subgroups with the indicated course profile.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

TABLE 4

Subgroup Differences on Individual Differences Dimensions and Selected Criteria—Females

Group	Subgroup characteristics in terms of the biographical factors and the test battery[a]	1968 n	Subgroup differences on criteria[b]					Subgroup characteristics in terms of course profile
			$\% \geq 1$ dean's list	$\% \geq 1$ probation or dismissal	$\% \geq 1$ course withdraws	Grade point average	% Arts–science	
1	High Achievement, high Warmth, low Status, low Athletic, low Science–Art	73	40	14–low	44	2.87	50	21% social science (vs. 13%)[c]
2	Low Reading, high Warmth, low Achievement ** high F scale, high Economic Values, low SAT	65	28	34	34	2.44	16–low	46% fewer hours group (vs. 17%)[c]
3	High Achievement, low Status, high Athletic, low Reading ** low Complexity, low Radicalism	59	44	15–low	31	2.89	38	12% business (vs. 21%)[c]
4	High Leader, low Academic Attitude, low Science–Art ** high F scale, high Complexity	38	58	24	39	2.80	29–low	32% education (vs. 21%)[c]
5	Low Popularity, low Maturity, low Leader ** low Complexity	47	39	26	30	2.79	46	25% fewer hours group
6	High Female Role, high Maladjustment, low Achievement	41	17–low	56–high	41	2.42–low	29–low	29% fewer hours group
7	Low Popularity, low Female Role, low Leader, high Science–Art ** low Complexity	56	38	35	31	2.71	66–high	18% science (vs. 5%)[c]
8	High Reading, low Dad's Girl, low Negative Emotions	41	27	37	41	2.44	36	20% home economics (vs. 11%)[c]
9	High Reading, high Negative Emotions, low Academic Attitude, low Female Role ** high Complexity, low F scale	22	19–low	24	48	2.59	38	
10	Low Maturity, high Negative Emotions, low Status ** high Simplicity, low SAT	35	15–low	59–high	41	2.34–low	25	37% fewer hours group
11	Low Academic Attitude, low Popularity	42	50	21	36	2.79	56	23% fine arts (vs. 14%)[c]
12	High Maladjustment, high Academic Attitude, low Popularity, high Science–Art ** high SAT	20	60–high	10–low	45	3.08–high	73–high	25% science
13	High Status, high Achievement ** high SAT, high Complexity	70	65–high	20	48	3.00–high	49	22% social science
14	Low Warmth, high Control, high Academic Attitude, low Science–Art	15	40	13–low	33	2.86	33	
15	High Leader, high Female Role, low Control, low Reading ** high F scale, low SAT	43	40	23	33	2.78	22–low	32% education
Average			28%	27%	38%	2.72	57%	
χ^2 or F			58.00***	57.20***	10.84	6.68***	31.14**	

[a] Differences on the biographical inventory are listed first. Differences on the test battery are preceded by two asterisks (**).
[b] Groups deviating substantially from the average value in an upward direction are marked high, those deviating in a downward direction are marked low.
[c] Percent of females in all subgroups with the indicated course profile.
* $p < .05$.

Perhaps more important than the statistical evidence was the fact that the relationships make reasonable sense. For example, male Group 1 had a high proportion of fine arts majors, and was characterized as being low in athletic interest, low in verbal aggression, and high on the Music Teacher scale of the SVIB. Male Group 3 was low in verbal aggression, low on the SATs, and low in complexity and had an extremely high proportion of business majors. Obviously, there are more features of interest than can be detailed in this discussion. Comparing the tables, especially the base rates, indicated the tremendous sex differences in the popularity of the majors, as well as differences resulting from the number of course hours.

Cross-validation

One question that remains concerns whether differences observed among biographical data groups in one year will hold up in succeeding years. The predicted grade point average and its components, high school grades and the SAT scores, were used for this aspect of the investigation. Cross-validation was studied by comparing the rank order of groups evolved in 1968 on predicted grade point average with rank of the 1970 groups on this same variable. In addition, the known 1968 groups and scores on three variables (high school grade point average plus the SAT scores) allowed for canonical discriminant analyses, one for each sex.

It is not necessary to list predicted grade point average means for the 1968 groups since they are similar to the actual grade point average means shown on Tables 3 and 4. The rank-order correlations between predicted and actual grade point averages for the 1968 groups was .85 for the males and .84 for the females, highly significant ($p < .001$) in both cases. The rank-order correlations between group means on predicted grade point average in 1968 and 1970 was .89 ($p < .001$) for the males and .73 ($p < .01$) for the females (both significant), indicating that the criterion differences were maintained over the two-year period.

The results of the discriminant analyses are summarized and discussed in detail (including plots of the 1968 and 1970 groups in the discriminant space) elsewhere (Schoenfeldt, 1970). Two discriminant functions accounted for 78% of the discriminating variance for the males and 73% for the females. For each sex, the first function was a high school grade point average dimension and the second an SAT total dimension. The two functions resulted in plots with four quadrants with the origin representing average grade point average and SAT scores. Four of the male groups formed in 1970, Groups 5, 8, 9, and 11 were in different quadrants from the corresponding 1968 groups. For the females, all but one of the 1970 groups was in the same quadrant as the 1968 counterpart. Group 5 was below the origin on the high school grade point average dimension in 1968 and above in 1970. On the basis of these results it can be concluded that criterion relationships do cross-validate.

DISCUSSION

These several validity studies indicate that the assessment–classification model is viable. The subgroups formed to be homogeneous with respect to antecedent behavior differed on univariate criteria related to job success. These groups were also related to job family subgroups, that is, groups formed due to similarities with respect to the path walked in working toward a baccalaureate degree. Finally, the subgroup differences observed in 1968 held up in a new sample surveyed in 1970.

What are the implications of this model on the profession and society? What is the role of the psychology of human utilization in the seventies and beyond? The assessment–classification model might serve the major perturbations resulting from the governmental regulations regarding validation. One trend the regulations seek to redress is the emphasis placed on an applicant's credentials, specifically the number and character of his degrees, rather than on his personal competence (Dailey, 1971). Indeed, a major thrust of the regulations is a shift away from the selection model that has dominated industrial psychology since its inception to a realization that most people *can* do many (if not most) of the jobs that need doing. Towards this end, the assessment-classification procedure should be in harmony with the regulations. It is a model aimed at matching individual tendencies (especially nonintellective) with the job requirements, and assumes both are competently measured. To the extent that there might be an excess of people to jobs, selection would be involved, although it is envisioned that application of the model might extend to many employers in a region, thus emphasizing classification.

Another realization made explicit by the regulations is the fact that personnel policy involves more than the hire decision. The principal transition points into the world of work are upon completion of secondary school or some form of higher education. The model is directly applicable to these entry levels. Beyond this, the model can be used for persons already employed, as, for example, in conjunction with the assessment centers which have received much attention during the past several years (Bray & Grant, 1966; Byham, 1971).

It is probably only a matter of time until the regulations will be extended to cover those experiences which qualify individuals for particular careers, such as technical training, college, graduate school, etc. The assessment–classification model is concerned with matching persons seeking opportunities, whether jobs, or other types of experiences, with available openings. Utilization of the model is not limited to traditional employment settings.

The complete list of benefits anticipated from the assessment–classification system is too lengthy to describe in detail. In summary, the model's significance is twofold. First, the model provides a new approach to the utilization of talent, an approach that promises to improve the fit of individuals and jobs and at the same time be in concert with the major thrusts of the governmental regulations regarding validation. Second, the model allows for the continuous collection of data of the kind needed to anticipate and study personnel problems before they become critical. When applied to a large employer or regionally, for example, it would provide feedback regarding shifts in applicant characteristics or job requirements.

REFERENCES

Anderson, B. B. An interinstitutional comparison on dimensions of student development: A step toward the goal of a comprehensive developmental-integrative model of human behavior. (Doctoral dissertation, University of Georgia) Ann Arbor, Mich.: University Microfilms, 1973. No. 73-5638.

Barrett, G. V. Research models of the future for industrial and organizational psychology. *Personnel Psychology*, 1972, **25**, 1–17.

Boardman, W. K., Calhoun, L. G., & Schiel, J. H., Jr. Life experience patterns and development of college leadership roles. *Psychological Reports*, 1972, **31**, 333–334.

Bray, D. W., & Grant, D. L. The assessment center in the measurement of potential for business management. *Psychological Monographs*, 1966, **80** (17, Whole No. 625).

Byham, W. C. (Chm.) Validity of assessment centers. Symposium presented at the meeting of the American Psychological Association, Washington, D.C. September 1971.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., III, & Weick, K. E., Jr. *Managerial behavior, performance, and effectiveness*. New York: McGraw-Hill, 1970.

Cirino Gerena, G. Toward a model for test prediction. *American Psychologist*, 1972, **27**, 759–760.

Claudy, J. G. An empirical investigation of small sample multiple regression and cross-validation. (Doctoral dissertation, University of Tennessee) Ann Arbor, Mich.: University Microfilms, 1970. No. 70-7553.

Cole, M., & Bruner, J. S. Cultural differences and inferences about psychological processes. *American Psychologist*, 1971, **26**, 867–876.

Cooley, W. W., & Lohnes, P. R. *Multivariate data analysis*. New York: Wiley, 1971.

Cunningham, J. W., Tuttle, T. C., Floyd, J. R., & Bates, J. A. *Occupational analysis inventory*. Raleigh, N.C.: Center for Occupational Education, 1970.

Dailey, C. A. *Assessment of lives*. San Francisco: Jossey-Bass, 1971.

Dunnette, M. D. A modified model for test validation and selection research. *Journal of Applied Psychology*, 1963, **17**, 317–323.

Fleishman, E. A. Development of a behavior taxonomy for describing human tasks: A correlational-experimental approach. *Journal of Applied Psychology*, 1967, **51**, 1–10.

Ghiselli, E. E. Differentiation of individuals in terms of their predictability. *Journal of Applied Psychology*, 1956, **40**, 374–377.

Ghiselli, E. E. The prediction of predictability. *Educational and Psychological Measurement*, 1960, **20**, 3–8.

Klein, H. A. Personality characteristics of discrepant academic achievers. (Doctoral dissertation, University of Georgia) Ann Arbor, Mich.: University Microfilms, 1973. No. 73-1514.

Kundert, K. R. Tools of analysis for pattern recognition. (Doctoral Dissertation, University of Georgia) Ann Arbor, Mich.: University Microfilms, 1973. No. 73-5725.

Lee, M. C. Interactions, configurations, and nonadditive models. *Educational and Psychological Measurement*, 1961, **21**, 797–805.

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. *The development and background of the Position Analysis Questionnaire (PAQ)*. West Lafayette, Ind.: Purdue University, Occupational Research Center, June 1969.

Meyer, H. H. The future for industrial and organizational psychology: Oblivion or millennium? *American Psychologist*, 1972, **27**, 608–614.

Novick, M. R., & Jackson, P. H. Bayesian guidance technology. *Review of Educational Research*, 1970, **40**, 459–494.

Owens, W. A. Toward one discipline of scientific psychology. *American Psychologist*, 1968, **23**, 782–785.

Owens, W. A. President's message. *The Industrial Psychologist*, 1970, 7(3), 1–2.

Owens, W. A. A quasi actuarial prospect for individual assessment. *American Psychologist*, 1971, **26**, 992–999.

Prien, E. P., & Ronan, W. W. Job analysis: A review of research findings. *Personnel Psychology*, 1971, **24**, 371–396.

Schoenfeldt, L. F. Life experience as a moderator in the prediction of educational criteria. Paper read at the meeting of the American Educational Research Association, Minneapolis, March 1970.

Taylor, L. R. *The State Farm job analysis project: Report of pilot study results*. Bloomington, Ill.: State Farm Insurance Companies, 1970.

Toops, H. A. The use of addends in experimental control, social census and managerial research. *Psychological Bulletin*, 1948, **45**, 41–74.

Toops, H. A. A research utopia in industrial psychology. *Personnel Psychology*, 1959, **12**, 189–225.

Vale, J. R., & Vale, C. A. Individual differences and general laws in psychology. *American Psychologist*, 1969, **24**, 1093–1108.

Ward, J. H., Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 1963, **58**, 236–244.

Ward, J. H., Jr., & Hook, M. E. Application of an hierarchical grouping procedure to a problem of grouping profiles. *Educational and Psychological Measurement*, 1963, **23**, 69–81.

Zedeck, S. Problems with the use of "moderator" variables. *Psychological Bulletin*, 1971, **71**, 295–310.

(Received November 1, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 5, 596-602

FARM AND FACTORY IN THE KIBBUTZ:
A STUDY IN AGRICO-INDUSTRIAL PSYCHOLOGY[1]

DOV EDEN
Department of Labor Studies,
Tel Aviv University, Israel

URI LEVIATAN[2]
Center for Social Research on the Kibbutz,
Givat Haviva, Israel

Kibbutz communities with a mix of factory and farm jobs were surveyed to study the effects of industrialization on work and workers while controlling for the effects of urbanization. Questionnaire responses of 476 workers in a representative sample of factories were compared to those of 175 workers in location-matched agricultural branches in 27 kibbutz collective settlements. Factory workers rated their jobs significantly lower than did farm workers on plant manager's leadership, opportunities for self-realization on the job, participation, control, peer relations, and information about the job. However, factory jobs were rated cleaner, easier, and more mental. Factory and farm workers were similar in outcome measures, including supervisory ratings of performance, job satisfaction, mental health, and alienation. Implications for kibbutz and nonkibbutz work are discussed.

The authors have never heard of the term *agricultural psychology*. At the time that Münsterberg's (1913) book, considered to be the first dealing with industrial psychology, was published, about 30% of the United States work force were still employed in agriculture (Taylor & Jones, 1964). However, farming appears to have largely missed out on the rise of scientific psychology. A recent review of *Psychological Abstracts* revealed only seven articles in American psychological journals between 1945 and 1970 that dealt directly with any aspect of agriculture (Richards, 1973). As the last half century has been characterized by a secular decline in the proportion of farmers in the United States work force (Taylor & Jones, 1964), it is easy to understand why the United States studies of the psychology of work have ignored the farmer. However, although industrial psychology has thrived for so long without the farmer, the study of work could be advanced by broadening its scope to include a type of work hitherto neglected. In developing nations, based solely on the number of workers involved, *agricultural psychology* would be relevant to the lives of many more workers than would industrial psychology. In addition, cross-occupational comparisons could be interesting and enlightening with regard to job content and employee attitudes.

In Israel, kibbutz[3] communities have been undergoing rapid industrialization. At present many of these so-called "collective farms" have a mix of industry and agriculture in which industry dominates the farm branches in total revenue, profits, number of workers, and in setting economic and social priorities in the settlement (Leviatan, 1973). Indeed, the kibbutz movement as a whole is ripe for agrico-industrial psychology.

---

[1] Support for the project was provided by the American Council for Behavioral Sciences in the Kibbutz and by the Fritz Naphtalie Foundation. Michal Palgi and Moshe Bar-Semech assisted.

[2] Requests for reprints should be sent to Uri Leviatan, Center for Social Research on the Kibbutz, Givat Haviva, Israel.

Uri Leviatan submitted his thesis under the name Uri Levitan; Uri Levitan and Uri Leviatan are one and the same person.

[3] A kibbutz (plural kibbutzim) is a collective settlement governed in principle by a general assembly of all members. Each kibbutz operates as a single economic unit and pays its members no wages. Food, clothing, housing, social services, and a nominal discretionary allowance are distributed by communal institutions on an agreed, egalitarian basis. A typical kibbutz has from 300 to 450 members. In 1970 the kibbutz population numbered 84,900, of whom 27,200 were employed adults aged 14 and above, that is, 3.9% of all employed adult males in the country (State of Israel, Central Bureau of Statistics, 1971). Work is organized collectively, and economic branch managers are elected by the branch's workers or by the general assembly of all members. In 1970 there were 170 factories in 224 kibbutzim (Leviatan, 1973). For additional readings about the kibbutz, see Darin-Drabkin (1963) and Rosner (1972).

Notwithstanding the resultant jump in standards of living, industrialization is widely considered to have brought dehumanization, alienation, anomie, and the like to the worker (Hauser, 1963). The same results are often attributed to urbanization. Since industrialization and urbanization commonly go hand in hand (Kuznets, 1963), it has been difficult to assess accurately the extent to which the oft bemoaned condition of modern man in industrial, urban society should be attributed to urbanization, to industrialization, or to both. The industrialization of the kibbutz provides a rare opportunity to tease out empirically the effects of industrialization and urbanization. Kibbutzim are incorporated as autonomous settlements and may be regarded as rural communities. At the same time, Israel is one of the most highly urbanized societies in the world. Thus the partially industrialized kibbutz provides an opportunity to compare farm and factory work in small settlements in an urbanized society. Whether the environment of the working kibbutz member should be regarded as rural or urban is moot and not crucial to the present article, since such macroenvironmental effects are controlled by comparing farm and factory workers in the same kibbutzim. Thus, in the kibbutzim it is possible to assess the effects of industrialization while controlling urbanization. Income is controlled, since all members of a given kibbutz, farmers and factory workers alike, share equally the fruits of their labor.

The present study was part of a research program on kibbutz industrialization undertaken by the Center for Social Research on the Kibbutz. (Previous analyses of some of the data from part of the sample have been reported by Levitan, 1970). This study attempted to examine the effects of industrialization on the worker while controlling urbanization. It was hypothesized that if industrial work deserves its ill-repute in comparison to agricultural work, then scores on measures of individual well-being of workers in industry (e.g., mental health, job satisfaction, etc.) would be lower than those of their agricultural counterparts. Further, if there are similar scores on such psychological measures for farm and factory, then it may be concluded that industrial work per se does not bring ill effects upon workers. This study also compared the features of farm and factory work to gain some understanding of the conditions that may give rise to differences in psychological outcomes.

METHOD

Populations and Samples

A population of factories and a population of farm branches were sampled. The industrial population was defined as all industrial plants that, in 1970, (*a*) were members of the Kibbutz Industrial Association (KIA), (*b*) were owned by individual kibbutzim (and not by interkibbutz conglomerates), (*c*) employed at least five workers, of whom no more than 30% were hired laborers (i.e., nonmember wage earners),[4] and (*d*) had been established at least one year prior to the study. Of the 157 plants affiliated with the KIA in 1970, 97 met the other criteria for selection. Ten of these were barred from the sample, since they had participated in a somewhat similar study only a year earlier (Rosner, 1971; Tannenbaum, Kavcic, Rosner, Vianello, & Wiesner, 1974). A three-dimensional classification of the remaining 87 plants was created using information about their movement affiliation,[5] size (small: 5–17 workers; medium: 18–30 workers; large: 31 or more workers), and technological level (mostly manual; semiautomated without feedback about deviation from standard operation; and semiautomated with feedback about deviation but without self-adjustment, as in Turner & Lawrence, 1965; fully automated plants did not exist in kibbutz industry at the time of the study). The sampling design called for choosing one plant at random from each of the 27 cells in the resulting $3 \times 3 \times 3$ table. The distribution of the 87 plants in the table was such that eight cells had zero entries. The original sampling design was therefore compromised by either randomly selecting a plant from the same movement and of the same size, or, if there was none, then randomly selecting a plant from the same movement. Thus, the sample of 27 factories can be regarded as fairly representative of the 87-member population of factories. The plants sampled manufactured metal, plastic, wood (furniture), food, and electronic products.

---

[4] Plants with more than 30% hired labor were eliminated so that the kibbutz factory in its purer form, in correspondence with the normative model for kibbutz enterprises, would be studied. This restriction of the population excluded that minority of the larger plants that employed most of the hired labor in the kibbutz industries.

[5] There are three major interkibbutz federations, known as movements. They differ in ideological coloring and in the kind and degree of normative organizational structure they advocate for their productive enterprises.

After the sampling of the industrial plants, one or more farm branches were selected from each kibbutz whose plant had been sampled. If the kibbutz had a poultry branch employing at least three workers (17 did), the poultry branch was selected first. This was done because the poultry branches were considered the most mechanized of the kibbutz agricultural branches, and it was regarded as important to assure their representation in the farm sample. In addition, 16 other agricultural branches were chosen in the same kibbutzim. These were mainly dairies, plantations, and field crop branches. Thus, the farm branches were chosen chiefly to serve as a useful comparison group to the industrial plants and not to represent the population of farm branches. The total number of members working in the 27 plants was 632, and in the 33 farm branches it was 224.

It has been claimed that advanced mechanization and application of high technology to modern farming, and to kibbutz farm branches in particular, have brought about the industrialization of agriculture and that a comparison of factory and farm workers would therefore really be a comparison of two types of industrial workers. For purposes of the present study, the definition of factory and farm work was made in terms of the distinction between primary and secondary economic activity. Those considered agricultural workers extract products from the earth, the sea, livestock, or man-made ponds and greenhouses. Those classified as factory, or industrial workers engage in manufacturing which transforms raw materials into more or less finished products.

Measures

A questionnaire with items covering a wide variety of organizational and social psychological variables was constructed and pretested. Many of the items had been used at the University of Michigan's Institute for Social Research and were translated into Hebrew and adapted for kibbutz use prior to pretesting. Other measures were created specifically for kibbutz use. Multiple-item indices were constructed by computing each respondent's mean score on the several items included in each index. Reliability was estimated by Cronbach's alpha (Nunnally, 1967). It was regarded as possible that the items might not hang together in the same way for industrial and farm workers, rendering their comparison meaningless. Therefore the inter-item correlations and the reliabilities for each index were computed separately for factory and farm workers. There were only minor differences between the two categories of workers in these intercorrelations and reliabilities, and the indices may be regarded as equivalently reliable for both samples. The reliabilities shown in Table 1 are based on data from all respondents in both samples together. The indices listed in Table 1 are ordered according to Likert's (1967) distinction between causal, intervening and end-result variables. With the sole exception of supervisors' ratings of workers, all measures were based on responses given by all workers who completed the questionnaire.

The five indices which were labeled causal variables included the following.

1. *Plant manager's leadership* index which combined items that assessed the extent to which the plant or branch manager was aware of workers' everyday job problems, was open to technical and organizational innovations, showed adaptability to change, and knew his job in general;
2. *Consideration* or the extent to which management was really interested in the workers' welfare and improved working conditions even when not forced to do so;
3. *Supervisory leadership* which was assessed by worker ratings of how frequently the immediate supervisor asked the workers' opinions, how good he was in relating to his subordinates, how friendly and approachable he was, and his success in coordinating his work group's interests with those of other groups in the plant;
4. *Opportunities for self-realization* which measured the extent to which the job provided opportunities for learning new things, using one's own ideas, doing interesting work, using one's abilities and knowledge, and trying unconventional work methods;
5. *Participation* which gauged workers' feelings regarding what part they took in making important decisions concerning their own work and general problems in the plant; also included in this index were workers' assessments of the actual authority of the worker assembly, a kibbutz institution designed to provide a forum for worker participation in managing the plants.

The six indices considered intervening variables included the following.

1. *Control* index or the extent to which the job gave the worker authority over other people, the degree of responsibility the worker had in the plant, and the degree of influence each worker felt he had personally;
2. *Peer relations* index which was comprised of items tapping group spirit and mutual aid, peer understanding of one's job problems, preferring present team to any other, peer work facilitation (Bowers & Seashore, 1966), and an overall rating of peer relations in the plant;
3. *Information* index which was constructed by averaging ratings of the extent to which each worker was "in the know" in his plant concerning the production plan, the investment plan, economic performance, and technical problems

involved in production, and a question asking how well the worker felt he knew how his job fit into the overall operation of the plant;

4. *Superior-subordinate communication* which was assessed by means of four items that measured the extent to which both the superior and the subordinate obtained and provided information needed for job performance;

5. *Responsibility* or the degree of responsibility felt for one's own work, for one's work group, and for the entire plant;

6. *Initiative* index which included measures of self-reported pursuit of better and faster work methods, doing extra work beyond that required, and trying out innovations on the job.

Four indices measured the end-results variables. The first three of these were constructed from items in the questionnaire. The four included:

1. *Job Satisfaction* index which was comprised of a measure of overall satisfaction with the plant and a measure of overall satisfaction with "the type of work you do";
2. Six-item *Mental health* index which included a pair of questions for each of three dimensions: depression, resentment, and self-esteem;
3. *Alienation* which, following Seeman's (1959) five-factor conceptualization, was quantified by means of a pair of items to measure each dimension: isolation, powerlessness, meaninglessness, self-estrangement, and normlessness. The 10 items were combined into a single index;
4. *Supervisory Ratings* (rather than self-report questionnaire items) were used to assess workers' performance. Supervisors were presented with lists of their subordinates and were asked to rate them on three 7-point scales in terms of their skills and knowledge, work quality, and dependability.

No *harder* measure of individual performance on which these workers could be compared was available, due to the diversity of their jobs. However, the validity of these ratings was established at the aggregate level by correlating the plant mean supervisory rating with the plant's capital recovery rate. This correlation was .35 ($df = 26$, $p < .05$), indicating some degree of validity for the supervisors' ratings.

In order to gauge the workers' perceptions of various aspects of the work itself, they rated nine characteristics of their jobs on pretested 7-point semantic differential scales. The bipolar labels attached to these scales appear in Figure 1.

Procedure

The top manager of each site sampled was informed about the research project in advance in a letter from the KIA and the Center for Social Research on the Kibbutz, soliciting his cooperation. One of the 27 sampled plants declined to participate since it was in the midst of a reorganization and expansion program. A replacement plant was chosen at random from the same movement. Field workers from the Center arrived at each plant on a prearranged date and appeared before the worker assembly to describe the research and administer the questionnaires. This took a total of about one hour at each plant, on company time. Absent workers received envelopes for mailing back their completed questionnaires directly to the Center. All questionnaires were anonymous. The procedure with the farm sample was similar except that group administration was not feasible due to the geographic dispersion of agriculturists during working hours. Therefore, they mostly were given the questionnaire individually with mail-back instructions. In all, 476 industrial and 175 farm workers returned usable questionnaires for a response rate of 76%.

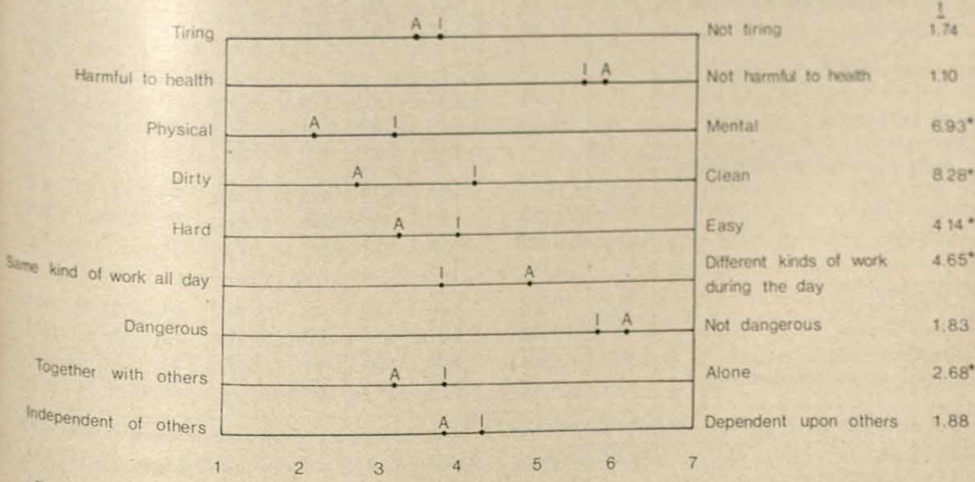


FIGURE 1. Mean scores for industrial (I) and agricultural (A) workers on semantic differential scales describing dimensions of jobs (* $p < .01$).

Farm managers were harder than their industrial counterparts to locate for the interview, resulting in a somewhat lower proportion of supervisory ratings for farm workers. All data were collected during the spring of 1970.

RESULTS

The comparison of factory and farm workers' scores on indices measuring causal, intervening, and end-result variables is shown in Table 1.

An analysis of the causal variables indicated that agricultural branch managers were rated significantly ($p < .01$) higher by their subordinates than were industrial plant managers by theirs. Nevertheless there was no difference in management consideration of farm and factory workers. The third row of Table 1 shows no difference in workers' ratings of their immediate supervisors' leadership. The two remaining causal indices showed that industry provided workers with significantly fewer opportunities for self-realization ($p < .05$) and for participation ($p < .01$) in decision making on their jobs.

Consistent with the report of less participation in industry was the significantly ($p < .01$) lower feeling of control experienced by factory workers. The differences in participation and control are the largest ones in Table 1. An examination of the other intervening variables revealed that peer relations and information were also significantly ($p < .01$) lower in industry. Vertical communication between workers and their immediate superiors was similar in agriculture and industry. Industrial workers considered themselves significantly ($p < .02$) less responsible than did the farm workers. Factory and farm workers exhibited equivalent levels of initiative-taking behavior.

The bottom third of Table 1 shows that factory and farm workers were strikingly

TABLE 1

RELIABILITY ESTIMATES AND MEAN SCORES FOR INDUSTRIAL AND AGRICULTURAL WORKERS ON CAUSAL, INTERVENING AND END-RESULT VARIABLES

Index	No. items	Coefficient alpha	Industry			Agriculture			
			M	*SD*	*n*	*M*	*SD*	*n*	*t*
Causal variables									
Plant manager's leadership	4	.73	4.11	.74	430	4.35	.62	126	3.30***
Consideration	2	.63	3.66	.93	449	3.56	1.09	164	1.09
Supervisory leadership	4	.74	3.74	.71	415	3.84	.71	146	1.48
Opportunities for self-realization	5	.84	3.01	.99	462	3.20	.81	173	2.18*
Participation	9	.88	2.74	.93	387	3.35	1.17	147	6.30***
Intervening variables									
Control	3	.74	3.12	1.12	472	3.71	1.00	171	6.09***
Peer relations	5	.73	3.30	.60	452	3.52	.70	171	3.91***
Information	5	.86	3.10	1.01	466	3.65	1.02	168	6.01***
Superior-subordinate communication	4	.81	4.09	.79	443	4.16	.70	132	<1
Responsibility	3	.65	3.98	.69	407	4.14	.63	138	2.38**
Initiative	3	.50	3.65	1.65	468	3.60	1.47	175	<1
End-result variables									
Job satisfaction	2	.76	3.60	.88	446	3.59	.93	171	<1
Mental health	6	.78	4.37	.86	451	4.41	.84	167	<1
Alienation	10	.83	3.82	.82	462	3.85	.81	167	<1
Supervisors' ratings of workers	3	.80	5.26	1.19	558[a]	5.37	1.35	130[b]	<1

[a] This *n* is larger than the self-report industrial sample, since plant managers filled out ratings on all their workers, some of whom did not subsequently return completed questionnaires.
[b] This *n* is smaller than the self-report agricultural sample, since managers of six of the participating farm branches were not available for interviewing.
* $p < .05$, two-tailed test.
** $p < .02$, two-tailed test.
*** $p < .01$, two-tailed test.

similar on psychological end-result variables, including job satisfaction, mental health, and alienation. The last row of Table 1 shows that supervisory ratings of workers were equivalently high for factory and farm workers.

Figure 1 shows that factory and farm jobs are indeed perceived differently by workers. Of the nine work characteristics measured by means of semantic differential scales, five showed highly significant differences ($p < .01$; two-tailed t tests based on responses from 500 industrial workers and 180 agricultural workers; the n varying slightly from scale to scale due to occasional missing data). Industrial work was considered more mental, cleaner, and easier, but also less varied in activity and more *alone*.

DISCUSSION

The semantic differentials used to measure the work itself offer strong evidence that there are indeed differences between work experiences in the kibbutz farm and factory. These data are sufficient to cast serious doubt upon the commonplace contention that the "industrialization of agriculture" in the kibbutz is so advanced that the kibbutz agriculturist has in effect become an industrial worker. Even if one accepts this view, the differences in Figure 1 indicate that the character of the work performed by this new industrial farmer is still quite different from that done by the industrial worker in the factory, and in any case it is worthwhile to study the differences between the two. It is perhaps not at all surprising that industrial work is perceived as less physical, cleaner, and easier than farm work. These are hygiene factors according to Herzberg, Mausner, and Snyderman (1959). These differences have a great deal of face validity and corroborate observations of the investigators during numerous site visits.

It is interesting that factory workers feel that their work is performed more *alone*. The average number of workers in the sampled factories and agricultural branches was 26 and 8, respectively. Thus, the factory workers actually reported themselves as being lonely in a crowd. Farm workers reported a greater feeling of working together with others, despite the smaller size and wider geographical dispersion of their branches.

Table 1 shows that, in addition to differences in the work itself, there are large and consistent differences in causal and intervening variables showing results unfavorable to the factory. The only variables showing no differences involve the immediate supervisors —their rating by their subordinates, their rating of their subordinates, and worker–supervisor communication. These similarities in items tapping aspects of worker–supervisor relations may be due to the common kibbutz milieu shared by members at different levels in the work organization. After hours they all remain a part of the same community, and nonwork interaction may serve to attenuate whatever differential effects might otherwise have occurred in interpersonal relations.

The causal and intervening variables showed overwhelming disadvantages for factory workers, and it is surprising that this did not produce differences in end results. Since the number of workers in the immediate work environment is negatively associated with job satisfaction (Indik, 1963), the size difference is an additional reason to expect industrial workers to be less satisfied. One could argue that alienation and mental health are more influenced by nonwork, and since kibbutz farm and factory workers share the same communities and nonwork social environment, they should be equally mentally healthy and alienated. This reasoning does not hold for job satisfaction, which should be influenced by job content and by causal and intervening variables. Why are the psychological returns to factory workers equivalent to those of farmers despite the evident disadvantages of factory work? On the basis of the data available one can only speculate as to the answer to this question.

The most obvious and straightforward interpretation of the results is that industrialization in and of itself has a measurable impact on the job itself and on the work environment, but not on the worker's psychological responses to his job. Among kibbutz members, similarity in degree of urbanization and comparability of all other community environment factors may attenuate or indeed obliterate the negative psychological impact of factory work. If this interpretation is true, one should seek the sources of the malaise of modern industrial man somewhere beyond the workplace. Future research could seek to reveal the mechanisms by which the kibbutz cushions its industrially employed members from the relative disadvantages of their work.

REFERENCES

Bowers, D. G., & Seashore, S. E. Predicting organizational effectiveness with a four-factor theory of leadership. *Administrative Science Quarterly*, 1966, **11**(2), 238–263.

Darin-Drabkin, H. *The other society*. New York: Harcourt, Brace & World, 1963.

Hauser, P. M. The social, economic, and technological problems of rapid urbanization. In B. F. Hoselitz & W. E. Moore (Eds.), *Industrialization and society*. The Hague: Mouton, 1963.

Herzberg, F., Mausner, B., & Snyderman, B. *The motivation to work*. New York: Wiley, 1959.

Indik, B. P. Some effects of organization size on member attitudes and behavior. *Human Relations*, 1963, **16**, 369–384.

Kuznets, S. Consumption, industrialization and urbanization. In B. F. Hoselitz & W. E. Moore (Eds.), *Industrialization and society*. The Hague: Mouton, 1963.

Leviatan, U. Process of industrialization in the Israeli kibbutzim. Paper presented at the Ninth International Congress of Anthropological and Ethnological Sciences, Chicago, August 1973.

Levitan, U. Status in human organization as a determinant of mental health and performance. Unpublished doctoral dissertation, University of Michigan, 1970.

Likert, R. *The human organization*. New York: McGraw-Hill, 1967.

Münsterberg, H. *The psychology of industrial efficiency*. Boston: Houghton Mifflin, 1913.

Nunnally, J. C. *Psychometric theory*. New York: McGraw-Hill, 1967.

Richards, J. M., Jr. The psychology of farming: A review of twenty-five years of research. *Journal of Vocational Behavior*, 1973, 3, 485–501.

Rosner, M. Hierarchy and democracy in kibbutz industry. Unpublished report. Givat Haviva, Israel: Center for Social Research on the Kibbutz, 1971 (Hebrew).

Rosner, M. Social aspects of industrialization in the kibbutz. In M. Rosner (Ed.), *The kibbutz as a way of life in modern society*. Givat Haviva, Israel: Center for Social Research on the Kibbutz, 1972.

Seeman, M. On the meaning of alienation. *American Sociological Review*, 1959, 24, 783–791.

State of Israel, Central Bureau of Statistics. *Statistical Abstract of Israel 1971*. Jerusalem: State of Israel, 1971.

Tannenbaum, A. S., Kavcic, B., Rosner, M., Vianello, M., & Wieser, G. *Hierarchy in organization*. San Francisco: Jossey-Bass, 1974.

Taylor, L., & Jones, A. R., Jr. *Rural life and urbanized society*. New York: Oxford University Press, 1964.

Turner, A. N., & Lawrence, P. R. *Industrial jobs and the worker: An investigation of response to task attributes*. Boston: Harvard University, Graduate School of Business Administration, 1965.

(Received September 17, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 5, 603–609

ORGANIZATIONAL COMMITMENT, JOB SATISFACTION, AND TURNOVER AMONG PSYCHIATRIC TECHNICIANS[1]

LYMAN W. PORTER,[2] RICHARD M. STEERS,[3] AND RICHARD T. MOWDAY
University of California, Irvine

PAUL V. BOULIAN
Yale University

This investigation studied changes across time in measures of organizational commitment and job satisfaction, as each related to subsequent turnover among a sample of recently employed psychiatric technician trainees. A longitudinal study across a $10\frac{1}{2}$-month period was conducted, with attitude measures collected at four points in time. The results of a discriminant analysis indicate that significant relationships existed between certain attitudes held by employees and turnover. Relationships between attitudes and turnover were found in the last two time periods only, suggesting that such relationships are strongest at points in time closest to when an individual leaves the organization. Organizational commitment was found to discriminate better between stayers and leavers than did the various components of job satisfaction.

The problem of employee turnover has continued to plague organizations in recent years despite an increase in investigations into factors affecting such behavior. With few exceptions, attitudinal studies of turnover have focused on the construct of job satisfaction as a predictor of tenure. A moderate and consistent relationship has generally been found across various samples between greater job satisfaction and the propensity to remain with the organization (Brayfield & Crockett, 1955; Herzberg, Mausner, Peterson, & Capwell, 1957; Porter & Steers, 1973; Vroom, 1964).

Recent investigations of turnover have typically attempted to replicate the findings of previous studies and have generally failed to look beyond the job satisfaction construct for other potential correlates of turnover. In addition, such research has been consistently static in nature, collecting attitudinal measures at one point in time and comparing such data with either present or future turnover rates. As pointed out by Lefkowitz and Katz (1969), attitudinal measures of factors associated with turnover do not remain constant over time. Because of this, it appears more appropriate to measure *patterns* of attitude changes over time as they relate to the propensity to leave than to rely on a single measurement at one point in time. Our knowledge of the turnover process could therefore be expanded, at least in part, by the following: (*a*) investigations into other potential constructs relating to turnover in addition to job satisfaction and (*b*) measures of attitudinal changes over time as they affect departure. The present study is designed to accomplish both of these objectives.

This study focuses on the attitude construct of organizational commitment, in addition to job satisfaction, and compares their respective predictive powers in differentiating stayers from leavers among a sample of psychiatric technician trainees. The assessment takes the form of a longitudinal study over a $10\frac{1}{2}$-month period.

Based on previous findings, we would expect overall job satisfaction measures to be inversely related to turnover (Hulin, 1966, 1968; Waters & Roach, 1971; Wild, 1970). When individual components of job satisfaction—satisfaction with pay, promotion, supervision, co-workers, and the work itself—have been studied, fairly consistent relationships also have been found with the propensity to remain (Farris, 1971; Friedlander & Walton, 1964; Hulin, 1968; Knowles, 1964; Ley, 1966). Such findings are not universal, however, as some studies found certain of these factors unrelated to withdrawal (Kilbridge, 1961; Taylor & Weiss, 1969b; Telly, French, & Scott, 1971; Waters & Roach, 1971).

---

[1] This research was carried out under a contract from the Office of Naval Research (Contract N00014-69-A-0200-9001 NR 151-315).

The authors wish to express their appreciation to Joseph Champoux, William Crampon, and Eugene Stone for their valuable assistance during various phases of this investigation.

[2] Requests for reprints should be sent to Lyman W. Porter, Graduate School of Administration, University of California, Irvine, California 92664.

[3] Now at the School of Economics and Management, Oakland University, Rochester, Michigan.

Organizational commitment may represent a related but more global evaluative linkage between the employee and the organization that includes job satisfaction among its specific components. Relatively few studies have examined the relationship between commitment and turnover behavior (Porter, Crampon, & Smith, 1972). *Organizational commitment* is defined in the present context in terms of the strength of an individual's identification with and involvement in a particular organization. Such commitment can generally be characterized by at least three factors: (*a*) a strong belief in and acceptance of the organization's goals and values; (*b*) a willingness to exert considerable effort on behalf of the organization; (*c*) a definite desire to maintain organizational membership. Some evidence exists that a stated intention to remain with the organization, a component of commitment, is strongly and inversely related to turnover (Atchison & Lefferts, 1972; Kraut, 1970). It is predicted, therefore, that individuals highly committed to an organization's goals and willing to devote a great deal of energy toward those ends would be inclined to remain with the organization in an effort to assist in the realization of such highly valued objectives.

Under certain circumstances, measures of organizational commitment may be more effective predictors of turnover than job satisfaction. For example, while the individual may be dissatisfied with either his or her pay or supervisor, a high degree of commitment to the organization and its goals may serve to override such dissatisfaction in the decision to continue participation in the organization. In other cases—for example, where money is very important to an employee and where he or she is highly dissatisfied with the salary—satisfaction with various aspects of the job may take precedence over commitment in the decision to participate.

The use of psychiatric technician trainees as subjects for such a study has several advantages. First, these trainees typically have high rates of turnover. Second, training requirements and procedures as well as the pay and promotional opportunities during the first year of employment are all standardized. (A more detailed examination of the job environment and role requirements of psychiatric technicians may be found in Rushing, 1964, and Strauss et al., 1964.) Finally, as increasing emphasis is placed on the health sciences, it is especially important to learn more about the job environment and employee attitudes of such workers.

METHOD

Subjects and Research Setting

The subjects in this study comprise two classes or groups of psychiatric technician trainees employed by a major west coast hospital for the mentally retarded. The training course for these employees was one year in duration, including 1,000 hours of supervised clinical experience and 400 hours of onsite classroom lectures. The clinical tasks required of the trainees were quite varied, ranging from feeding and dressing patients to doing physical therapy and using such techniques as behavior shaping and operant conditioning. During the training period, the technicians received a modest salary.

The demographic characteristics of the two training groups were very similar. There were no differences between the groups with respect to age, male–female ratio, or educational background. In addition, selection procedures for both groups were identical, as were the training sequences, ward assignments, and instructors. The training groups were begun two months apart. Because of the similarity of the two groups, it was felt combining them into one study group was justified for purposes of analysis.

As with most state hospitals, this organization had a long history of relatively high turnover among its psychiatric workers, with rates ranging between 30% and 50% during the first year of full-time employment (i.e., after completion of training) and averaging 25% thereafter. A total of 84 individuals began training in Groups 1 and 2. However, by the beginning of this study (approximately nine months later), only 60 individuals remained. These 60 individuals constituted the initial study group. During the remainder of the study period, however, 27 more individuals terminated, or 45% of the sample.

TABLE 1

MEAN ATTITUDE SCORES AND STANDARD DEVIATIONS FOR STAYERS AND LEAVERS ACROSS TIME PERIODS

Attitude	Time Period 1 (−10 weeks)				Time Period 2 (−2 weeks)				Time Period 3 (+2 weeks)				Time Period 4 (+6 weeks)			
	Stayer (*n* = 16)		Leaver (*n* = 13)		Stayer (*n* = 25)		Leaver (*n* = 21)		Stayer (*n* = 27)		Leaver (*n* = 21)		Stayer (*n* = 28)		Leaver (*n* = 17)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Organizational commitment	4.0	1.0	3.5	1.0	4.3	1.1	3.5	.91	4.3	.96	3.3	.88	4.0	1.1	3.0	.98
Job satisfaction (JDI)																
Overall	151.9	29.5	130.0	56.1	152.0	43.2	127.7	45.0	155.4	53.5	136.4	44.8	146.9	44.7	124.5	50.2
Supervision	31.2	11.8	24.0	17.0	32.9	14.3	26.4	12.5	32.3	15.4	27.8	13.2	36.6	10.9	29.1	15.8
Co-workers	42.3	10.6	37.4	12.1	35.7	13.0	32.1	15.7	39.3	12.6	37.6	12.8	40.0	13.2	35.2	17.6
Work	35.6	5.8	30.2	9.0	32.9	8.8	27.3	12.5	27.7	10.5	24.5	12.2	26.3	12.3	22.8	12.1
Pay	20.5	12.4	11.5	10.8	20.2	10.5	19.2	12.7	21.9	14.5	21.3	14.4	20.5	13.9	19.9	14.0
Promotion	22.3	14.9	26.9	18.2	30.4	18.0	24.0	15.5	34.1	17.7	22.2	16.5	23.5	16.0	17.5	14.4

Note. JDI = Job Descriptive Index.

Instruments

Two research instruments were used in the attitudinal assessment. Each is discussed briefly.

Organizational Commitment Questionnaire. This 15-item questionnaire was designed to measure the degree to which subjects feel committed to the employing organization. Included in this instrument are items pertaining to the subject's perceptions concerning his loyalty toward the organization, his willingness to exert a great deal of effort to achieve organizational goals, and his acceptance of the organization's values. All items represent statements to which the subject responds on 7-point Likert-type scales, ranging from "strongly disagree" to "strongly agree." The wording of six items is reversed in an attempt to reduce response set bias. A measure of overall commitment for each respondent is derived by taking the mean score across all items. The internal consistency of the instrument as measured by coefficient alpha (Cronbach, 1951) ranged from .82 to .93 across the four time periods.

Job Descriptive Index (JDI). The JDI was employed to measure satisfaction with five aspects of the subjects' jobs: supervision, co-workers, work, pay, and promotion. In addition, the unweighted sum of the five scales was used as a measure of overall satisfaction. Validity and reliability data, as well as scoring techniques, for the JDI are described in Smith, Kendall, and Hulin (1969).

Data Collection Procedures

Instruments were administered to the subjects during four sequential time periods by the researchers. Confidentiality of responses was assured by both the university researchers and hospital management. Moreover, subjects were given the opportunity to decline participation but none did so.

The first administration (designated Time Period 1) was given 10 weeks prior to the completion of the training program. Only Group 1 participated in this administration since Group 2 had already passed this point in their training program. The remaining three administrations (Time Periods 2, 3, and 4) were given to both groups two weeks prior to training completion, two weeks after beginning full-time work (i.e., after training), and six weeks after beginning full-time work. Subjects who did not complete the training program were dropped from the investigation. The sample size fluctuated somewhat during the latter three administrations due primarily to incomplete questionnaire responses during all time periods and to a very small amount of turnover between the third and fourth administration periods. Thus, while a pure longitudinal design was attempted, it was not possible to achieve an entirely homogeneous sample across time. Turnover data, based on hospital records, were collected up to eight months following the completion of the training program.

RESULTS

The sample was divided into stayers and leavers based on subsequent turnover data, and mean scores for each group were calculated on the organizational commitment and job satisfaction measures. Group means, standard deviations and sample sizes for each time period are presented in Table 1.

The homogeneity of stayers and leavers concerning demographic characteristics is reflected in the absence of significant differences between the two groups with respect to education, male–female ratio and income. However, the mean age for stayers ($\bar{X} = 31.9$ years) was significantly higher ($p < .01$) than the mean age for leavers ($\bar{X} = 23.9$). Relationships between age and turnover have been found elsewhere and were not surprising findings in the present study (Farris 1971; Fleishman & Berniger, 1960; Ley, 1966; Minor, 1958; Stone & Athelstan, 1969). This finding suggests though, the possibility that any relationships found between attitudes and turnover may be attributable to differences in age. Consequently, age was controlled in the subsequent analysis by partialing its influence from the commitment and job satisfaction scores using a method suggested by Nunnally (1967).

A discriminant analysis (Overall & Klett, 1972) was performed for each time period to find the linear combination of the commitment and job satisfaction scores (with the influence of age partialed out) that best discriminates between groups formed on the basis of turnover. The significance of the discriminant functions was tested using a test statistic which is approximately distributed as a chi-square variable when computed by the procedure in Overall and Klett (1972). In the two group cases of discriminant analysis, only one discriminant function is found, and the results are functionally equivalent to those that would be found from a multiple regression using a dummy criterion variable (Tatsuoka, 1971). The total discriminatory power of the discriminant function was calculated using Tatsuoka's (1970) multivariate extension of the estimated omega-squared. The omega-squared statistic can be interpreted as the percentage of variance in the discriminant space that is relevant to group differentiation (Tatsuoka, 1970).

The discriminant analyses examined the relationship between turnover and the following set of variables: commitment to the organization; satisfaction with supervision; satisfaction with co-workers; satisfaction with the work itself; satisfaction with pay; satisfaction with opportunities for promotion. The measure of overall job satisfaction was not considered in the discriminant analysis because it is derived from a summation of the five subscales and thus its use (in conjunction with the subscales) would violate the mathematical requirements of the technique.

Results of the discriminant analysis for each time period are reported in Table 2. As the results indicate, no significant discrimination was found between stayers and leavers in either Time Period 1 or Time Period 2. However, significant relationships were found between turnover and the several attitude measures in Time Period 3 ($p < .05$) and in Time Period 4 ($p < .05$).

The fact that significant relationships were found in the latter two time periods but not in Time Periods 1 and 2 suggests that the relationship between turnover and the several attitudes is strongest at points in time closest to when the individual terminates. In other words, as an individual approaches the point of leaving the organization it appears that his or her attitudes are a more accurate predictor of such subsequent behavior.

The results of the discriminant analysis for Time Period 3 indicate that 21% of the variance in commitment and job satisfaction is relevant to the subsequent decision to stay or leave the organization. Interpretation of the standardized discriminant weights (Table 2) allows conclusions to be drawn concerning the relative contribution of each variable in discriminating between groups. Commitment to the organization was clearly the most important variable in differentiating between stayers and leavers. Satisfaction with opportunities for promotion and satisfaction with the work itself were next most important. The moderate size of the discriminant weights for these variables relative to the weight for commitment indicates a smaller contribution to the discrimination between groups.

TABLE 2

STANDARD DISCRIMINANT WEIGHTS, TEST STATISTIC, AND DISCRIMINATORY POWER ACROSS TIME PERIODS

Variable	Time Period 1	Time Period 2	Time Period 3	Time Period 4
Standard discriminant weights				
Organizational commitment	−.12	1.04	1.04	1.43
Job satisfaction (JDI)				
Supervision	−.25	.05	−.24	−.12
Co-workers	.48	−.38	−.19	−.25
Work	.57	.10	−.50	−.39
Pay	.85	−.18	−.01	−.28
Promotion	−.40	.19	.52	.01
Test statistic	5.1	4.7	13.5*	13.0*
Degrees of freedom	6	6	6	6
Total discriminatory power (%)	12.5	7.4	20.7	21.0

Note. JDI = Job Descriptive Index.
* $p < .05$.

Examination of the group means for stayers and leavers on the original measures in Time Period 3 indicates that stayers consistently had more positive attitudes than did leavers. The only exception to this was found with respect to satisfaction with pay, where the mean for stayers and leavers was approximately equal.

Results of the discriminant analysis in Time Period 4 indicate that 21% of the variance in the discriminant space was relevant to turnover. Examination of the discriminant weights (Table 2) indicates that commitment to the organization was again the most important variable in the discrimination between groups. The size of the discriminant weights for the other variables relative to the size of the weight for commitment indicates that the other variables contributed little, if anything, to the relationship found. Group means on the original measure in the last time period indicate that stayers expressed higher levels of commitment and job satisfaction than did leavers. The only exception was satisfaction with pay, where stayers and leavers reported comparable levels of satisfaction.

DISCUSSION

The primary purpose of this investigation was to study patterns in organizational commitment and job satisfaction across time as they relate to turnover among a sample of employees—psychiatric technicians, in this case—typically having high rates of turnover. The results clearly indicate that the attitudes held by an individual are predictive of subsequent turnover behavior, with individuals who ultimately leave the organization having less favorable attitudes than individuals who stay. Patterns of attitudes across time suggest that this inverse relationship between favorable attitudes and turnover generally is stronger as individuals approach the point at which they leave the organization. Based on these results, several points warrant emphasis and discussion.

When group means on the original measures are viewed within a cross-sectional framework, it is apparent that the level of attitudes for leavers is consistently lower than the level of attitudes for stayers during all four time periods. This finding is particularly evident with respect to commitment to the organization. Although it is clear that leavers are characterized by lower levels of commitment and satisfaction, the data are not sufficient to demonstrate why such differences occur.

One possible explanation may be found in the extent to which the expectations of stayers and leavers had been met on the job (Porter & Steers, 1973). Each individual can be viewed as bringing a unique set of expectations to the employment situation. The relatively common reward system (i.e., levels of pay, training procedures, task requirement, and instructors and supervisors were the same for all subjects) can be regarded as approximating more closely the level of expectations of those individuals who remained with the organization than those who left. Within such a framework, certain attitudes and turnover would be viewed as outcomes of a process in which individuals compare their level of expectations with the perceived realities of the job environment. While the present investigation was not specifically designed to examine this issue, the importance of considering met expectations in the withdrawal process has been suggested elsewhere in the literature (Dunnette, Arvey, & Banas, 1973; Katzell, 1968; Macedonia, 1969; Weitz, 1956; Youngberg, 1963).

When the results are viewed longitudinally, the relationship between attitudes and turnover appears to grow generally stronger over time. Evidence that a general decline in the level of attitudes takes place prior to leaving the organization has been found elsewhere (Porter et al., 1972). Such a trend was especially evident in this study with respect to organizational commitment, where the difference between the stayer and leaver mean scores generally increased over time. However, such a relationship was not as evident on the various measures of satisfaction. In fact, the difference between the stayer and leaver means on the subscales of the JDI tended to *decrease* over time, although such a pattern was not entirely consistent. In addition, the mean satisfaction scores for stayers and leavers fluctuate considerably across time with no apparent pattern.

The differential relationships between turnover and the two attitude constructs suggests that commitment and satisfaction are related yet distinguishable attitudes. Across the four time periods, the average intercorrelation between commitment and the five subscales of the JDI were .50, .42, .40, and .55, respectively (the highest correlations were generally found between commitment and satisfaction with the work itself). Thus, on the average, the two constructs shared less than 35% of the common variance. When partial correlations were calculated between commitment and turnover holding level of overall satisfaction (JDI total) constant, significant relationships were found in three of the four time periods (in Time Period 1, neither commitment nor overall satisfaction was related to turnover). These findings suggest that although we would expect commitment and satisfaction to be related, each construct appears to contribute unique information about the individual's relationship to the organization.

Although this investigation was not designed to examine in any extensive manner the relationship between the two constructs, it is possible to speculate as to the nature of the differences between commitment and satisfaction. On the one hand, the development of organizational commitment appears to require an individual to think in fairly global terms about his or her relationship to the organization during the initial employment period. We would expect that a relatively greater amount of time would be required for an employee to determine his level of commitment to the organization than would be the case with his level of job satisfaction. This process is likely to result in a stable and enduring level of commitment. On the other hand, the degree of one's job satisfaction appears to be largely associated with specific and tangible aspects of the work environment and may represent a more rapidly formed affective response than does commitment. One's level of job satisfaction may thus be more transitory in nature, which could account for the shifts in the mean degrees of satisfaction over time found in the present study.

The overall findings of this study suggest that in this type of organizational context, general attitudes toward the organization may be more important in the decision to remain than the more specific attitudes toward one's particular job. It must be remembered, however, that such findings emerged among a sample of psychiatric technicians working in a state mental hospital. Given the nature and purpose of such an organization, it is conceivable that organizational commitment may take on a special importance in the participation decision. It is possible that some employees place a high value on the goals of the organization and such an orientation may outweigh—to a degree, at least—their like or dislike of the particular tasks required to attain such goals. Whether similar results would be found among employees of most profit-centered business organizations remains to be demonstrated.

REFERENCES

Atchison, T. J., & Lefferts, E. A. The prediction of turnover using Herzberg's job satisfaction technique. *Personnel Psychology*, 1972, **25**, 53–64.

Brayfield, A. H., & Crockett, W. H. Employee attitudes and employee performance. *Psychological Bulletin*, 1955, **52**, 396–424.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, **16**, 297–334.

Dunnette, M. D., Arvey, R., & Banas, P. Why do they leave? *Personnel*, 1973, **50**, 25–39.

Farris, G. F. A predictive study of turnover. *Personnel Psychology*, 1971, **24**, 311–328.

Fleishman, E. A., & Berniger, J. One way to reduce office turnover. *Personnel*, 1960, **37**, 63–69.

Friedlander, F., & Walton, E. Positive and negative motivations toward work. *Administrative Science Quarterly*, 1964, **9**, 194–207.

Herzberg, F., Mausner, B., Peterson, R. O., & Capwell, D. F. *Job attitudes: Review of research and opinion*. Pittsburgh: Psychological Service of Pittsburgh, 1957.

Hulin, C. L. Job satisfaction and turnover in a female clerical population. *Journal of Applied Psychology*, 1966, **50**, 280–285.

Hulin, C. L. Effects of changes in job-satisfaction levels on employee turnover. *Journal of Applied Psychology*, 1968, **52**, 122–126.

Katzell, M. D. Expectations and dropouts in schools of nursing. *Journal of Applied Psychology*, 1968, **52**, 154–157.

Kilbridge, M. Turnover, absence, and transfer rates as indicators of employee dissatisfaction with repetitive work. *Industrial and Labor Relations Review*, 1961, 15, 21–32.

Knowles, M. C. Personal and job factors affecting labor turnover. *Personnel Practice Bulletin*, 1964, **20**, 25–37.

Kraut, A. I. *The prediction of turnover by employee attitudes.* New York: IBM World Trade Corporation, 1970.

Lefkowitz, J., & Katz, M. Validity of exit interviews. *Personnel Psychology*, 1969, **22**, 445–455.

Ley, R. Labor turnover as a function of worker differences, work environment, and authoritarianism of foremen. *Journal of Applied Psychology*, 1966, **50**, 497–500.

Macedonia, R. M. Expectation-press and survival. Unpublished doctoral dissertation, New York University, Graduate School of Public Administration, 1969.

Minor, F. J. The prediction of turnover of clerical employees. *Personnel Psychology*, 1958, **11**, 393–402.

Nunnally, J. *Psychometric theory.* New York: McGraw-Hill, 1967.

Overall, J. E., & Klett, C. J. *Applied multivariate analysis.* New York: McGraw-Hill, 1972.

Porter, L. W., Crampon, W. J., & Smith, F. J. *Organizational commitment and managerial turnover: A longitudinal study.* (Tech. Rep. 13, Individual-Organization Linkages Research Project) Irvine, Calif.: University of California, 1972.

Porter, L. W., & Steers, R. M. Organizational, work and personal factors in employee turnover and absenteeism. *Psychological Bulletin*, 1973, **80**, 151–176.

Rushing, W. *The psychiatric professions.* Chapel Hill, N.C.: University of North Carolina Press, 1964.

Smith, P. C., Kendall, L. M., & Hulin, C. L. *The measurement of satisfaction in work and retirement: A strategy for the study of attitudes.* Chicago: Rand McNally, 1969.

Stone, T. H., & Athelstan, G. T. The SVIB for women and demographic variables in the prediction of occupational tenure. *Journal of Applied Psychology*, 1969, **53**, 408–412.

Strauss, A., Schatzman, L., Ruder, R., Erhlich, D., & Sabshin, M. *Psychiatric ideologies and institutions.* New York: Free Press of Glencoe, 1964.

Tatsuoka, M. M. Discriminant analysis. The study of group differences. Champaign, Ill.: Institute for Personality and Ability Testing, University of Illinois, 1970.

Tatsuoka, M. M. *Multivariate analysis.* New York: Wiley, 1971.

Taylor, K., & Weiss, D. *Prediction of individual job turnover from measured job satisfaction.* (Research Rep. 22, Work Adjustment Project) University of Minnesota, May 1969. (a)

Taylor, K., & Weiss, D. *Prediction of individual job termination from measured job satisfaction and biographical data.* (Research Rep. 30, Work Adjustment Project) University of Minnesota, October 1969. (b)

Telly, C. S., French, W. L., & Scott, W. G. The relationship of inequity to turnover among hourly workers. *Administrative Science Quarterly*, 1971, **16**, 164–172.

Vroom, V. *Work and motivation.* New York: Wiley, 1964.

Waters, L. K., & Roach, D. Relationship between job attitudes and two forms of withdrawal from the work situation. *Journal of Applied Psychology*, 1971, **55**, 92–94.

Weitz, J. Job expectancy and survival. *Journal of Applied Psychology*, 1956, **40**, 245–247.

Wild, R. Job needs, job satisfaction, and job behavior of women manual workers. *Journal of Applied Psychology*, 1970, **54**, 157–162.

Youngberg, C. F. An experimental study of job satisfaction and turnover in relation to job expectancies and self expectancies. Unpublished doctoral dissertation, New York University, 1963.

(Received October 26, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 5, 610–615

PREDICTING ABSENTEEISM AND TURNOVER: A FIELD COMPARISON OF FISHBEIN'S MODEL AND TRADITIONAL JOB ATTITUDE MEASURES

JOHN E. NEWMAN[1]

University of Illinois

The present investigation studied the lab-to-field generalizability of Fishbein's attitude–behavior model and examined the relative efficacy of the Fishbein model and traditional job attitude measures as predictors of absenteeism and turnover. Predictor data were collected from 108 nursing home employees immediately preceding the two-month time period of interest. Criterion data were obtained at the end of the two-month period. Fishbein's model received some field support, particularly with respect to predicting turnover. Traditional job attitude measures were more effective predictors of absenteeism, while Fishbein's model was a more effective predictor of turnover. It was concluded that neither approach seems superior especially in light of the amounts of criterion variance accounted for.

Empirical studies of the attitude–behavior relationship in specific reference to job attitudes and job withdrawal behaviors have been reviewed recently by Porter and Steers (1973). They concluded that there is considerable evidence for consistent negative relationships between job attitudes (i.e., job satisfaction) and turnover. There appears to be a similar relationship of job attitudes to absenteeism, although there is much less information available. Porter and Steers note that many of the more recent findings are based on reliable and valid measures of job attitudes such as the Job Descriptive Index, JDI, (Smith, Kendall, & Hulin, 1969).

Although these observed attitude–behavior relationships are relatively consistent, they are seldom overly strong. Obtained correlations or multiple correlations as high as .50 have been rare. Since the prediction of job withdrawal behavior is a critical research topic from both theoretical and applied perspectives, it is important to further specify the precise nature of the relationship between attitudinal measures and withdrawal behavior and to test these models in applied settings.

Fishbein (1967) has developed a model for predicting overt behavior from verbal predictors. Fishbein's theory is an adaptation of Dulany's (1961, 1962) theory of propositional control and a variation of the extended instrumentality theories of behavior (Mitchell & Biglan, 1971).

The theory identifies two kinds of variables that function as the basic determinants of behavior: (*a*) attitudes toward the behavior and (*b*) normative beliefs. The first component, attitudes toward the behavior, reflects the instrumentality concept and depends upon the individual's beliefs about the consequences of performing the particular behavior in a given situation and his evaluation of these consequences. The second component, normative beliefs, depends upon the individual's beliefs regarding others' expectations of his behavior and his motivation to comply with these expectations. Fishbein acknowledges that other variables may affect behavior, but suggests that they operate indirectly by influencing one of the above determinants.

Fishbein's theoretical formulation may be expressed in symbolic form as a multiple regression equation:

$$B \approx BI = (A_{act})\ w_0 + (NB_s \cdot MC)\ w_1$$

Where:

B = overt behavior
BI = behavioral intention

---

[1] The author wishes to thank Martin Fishbein, Charles Hulin, and the reviewers for their valuable comments on an earlier draft of this article.

Requests for reprints should be sent to John Newman who is now at One State Farm Plaza, State Farm Insurance Companies, Bloomington, Illinois 61701.

A_{act} = attitude toward the behavior in a given situation
NB_s = social normative beliefs, i.e., perceived expectations of others
MC = motivation to comply with the normative beliefs
w_0, w_1 = empirically determined weights.

It is important to note that the criterion predicted by the components of this model is the individual's behavioral intention (BI). Ajzen and Fishbein (1969) have stated that BI serves as an intervening variable between the attitudinal and normative predictors and the overt behavior. They believe that the best predictor of a given behavior should be the person's intention to engage in that behavior. According to the theory then, if one can predict BI, one can also predict B with only slight attenuation.

The results of more than nine laboratory studies have indicated that the average correlation between BI and a wide range of behaviors is about .70 (Fishbein, in press). These same studies have indicated that the average multiple correlation between the two components of Fishbein's model and behavioral intentions is about .80 (Fishbein, in press). The model, however, has not been tested in on-going work organizations, and the usefulness of the theory for organizational situations is unknown.

The purpose of this study then was to test Fishbein's model in a real organizational setting. In addition to exploring the generalizability of the model from the laboratory to the field, the relative efficacy of the Fishbein model and traditional job satisfaction measures as predictors of job withdrawal behavior was examined.

METHOD

Subjects and Research Setting

The subjects ($N = 108$) were employees of a county nursing home and included nurse's aides, licensed practical nurses, registered nurses, administrators, staff, food service and housekeeping personnel. Slightly over 88% were female.

Questionnaire

The questionnaire provided measures of the following variables:

1. BI with respect to unexcused absences was measured in the following manner:

I am going to be absent (unexcused) from work on one or more days within the next two months (July and August).

EXTREMELY LIKELY	QUITE LIKELY	SLIGHTLY LIKELY	NEITHER LIKELY NOR UNLIKELY	SLIGHTLY UNLIKELY	QUITE UNLIKELY	EXTREMELY UNLIKELY

This scale was scored from 1 (extremely unlikely) through 7 (extremely likely).

2. Attitude toward being absent (A_{act}) was measured by taking the sum over the following four semantic differential scales with high loadings on the evaluative factor (cf. Osgood, Suci, & Tannenbaum, 1957):

Being absent (unexcused) from work on one or more days within the next two months (July and August) is:

foolish ____: ____: ____: ____: ____: ____: ____ wise
good ____: ____: ____: ____: ____: ____: ____ bad
harmful ____: ____: ____: ____: ____: ____: ____ beneficial
rewarding ____: ____: ____: ____: ____: ____: ____ punishing

Each scale was scored 1 through 7, 7 representing the positive end of the continuum.

3. NB_s with respect to unexcused absenteeism was measured by the following scale. Responses were scored from 1 (extremely unlikely) through 7 (extremely likely). Note that the phrase "whose opinions I respect" is an attempt to incorporate motivation-to-comply into this measure.

Most of the people, *whose opinions I respect*, think I should be absent (unexcused) from work on one or more days within the next two months (July and August).

EXTREMELY LIKELY	QUITE LIKELY	SLIGHTLY LIKELY	NEITHER LIKELY NOR UNLIKELY	SLIGHTLY UNLIKELY	QUITE UNLIKELY	EXTREMELY UNLIKELY

4. BI, A_{act}, NB_s with respect to voluntary resignation were also measured according to the above three formats.

5. Attitude toward the job in general was measured by the General Motors Faces scale (Kunin, 1955). This was scored from 1 (very dissatisfied) through 5 (very satisfied). It incorporated five faces (numbers 1, 4, 6, 8, 11) of the original scale.

6. Attitudes toward specific aspects of the employee's job situation (A_s) were measured by the five scales of the Job Descriptive Index (JDI) (Smith, Kendall, & Hulin, 1969). These scales measure the employee's satisfaction with the work, supervision, promotion, pay, and co-workers.

Procedure

The predictor data were obtained from the participants in the field setting. In an effort to minimize the reactive nature of the study, the questionnaire was administered to small groups of employees. When the employee entered the questionnaire–administration room, he signed a numbered attendance roster and was given, unobtrusively as possible, a questionnaire with a corresponding number. This allowed matching of the questionnaire data with absenteeism and resignation data.

The general nature of the study was explained briefly, the participants were guaranteed confidentiality, informed it was not necessary to sign the questionnaire, and then asked to complete the questionnaire. The data collection occurred during each of the three work shifts on the two days immediately preceding the two-month time period of interest.

Data relevant to the behavioral criteria, unexcused absenteeism (a single dichotomous act criterion) and voluntary resignation (also a single dichotomous act criterion) were obtained from the organization's records at the end of the two-month period. Absenteeism was scored as 0 or 1 depending on whether the employee was (1) or was not (0) absent (unexcused) at all during the two-month period. It is very important to understand that we were predicting whether or not the employee would be absent, not frequency of absences as a continuous variable. Voluntary resignation was scored as 0 or 1 depending on whether the employee voluntarily resigned (1) or not (0) during the two-month period.

RESULTS

Fishbein's Model: Lab versus Field

The means, standard deviations, and intercorrelations of all predictors and both criteria are presented in Table 1. From Table 1 it can be seen that BI correlated .10 and .39 with absenteeism and resignation, respectively. Although the .39 correlation was significant ($p < .01$), it was substantially smaller than BI–B correlations ($\bar{r} = .70$) obtained in the laboratory studies mentioned earlier.

The Fishbein model was designed specifically to predict behavioral intentions. The multiple correlations between the respective attitudinal and normative components of the model and the intent to be absent ($R = .45$, $p < .01$) and with the intent to resign ($R = .70$, $p < .01$) indicate that Fishbein's model is a relatively effective predictor of behavioral intention. Both of these multiple correlations, however, were lower than those ($\bar{r} = .80$) obtained in the laboratory studies.

In summary, although the Fishbein model yielded several significant correlations, the amounts of variance in overt behaviors and in behavioral intentions accountable by the Fishbein model in this field test were substantially lower than those accountable by the model in laboratory situations (Table 1).

Predictive Power: Comparison of the Fishbein Model and the Traditional Job Attitude Measures

Absenteeism. BI correlated .10 with absenteeism (Table 1). A_{act} and NB_s had a multiple correlation of .12 with absenteeism (Table 2). Thus, Fishbein's model accounted for (or predicted) about 1% of the variance in this withdrawal behavior.

Examination of Table 1 indicates that a traditional measure of satisfaction with the work itself (JDI Work scale) correlated significantly with absenteeism ($r = -.19$). This correlation coefficient was larger than any achieved by Fishbein's variables. However, the best single predictor of absenteeism was satisfaction with the job-in-general (Faces scale, $r = -.31$), accounting for roughly 9% of the criterion variance.

TABLE 1

MEANS, STANDARD DEVIATIONS, AND INTERCORRELATIONS OF PREDICTORS AND CRITERIA

Measure	$\bar{X}$	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Absenteeism																
1. A_{act}	6.9	3.8	1.00													
2. NB_s	1.8	1.2	.36	1.00												
3. BI	1.6	1.3	.36	.38	1.00											
4. B	.1	.3	.02	.12	.10	1.00										
Voluntary resignation																
5. A_{act}	10.3	7.0	.48	.27	.25	−.03	1.00									
6. NB_s	2.0	1.6	.32	.40	.17	.16	.49	1.00								
7. BI	2.1	1.8	.38	.36	.35	.06	.66	.53	1.00							
8. B	.1	.3	.17	.24	.18	.23	.30	.32	.39	1.00						
JDI scales																
9. Work	31.6	9.6	−.33	−.32	−.23	−.19	−.16	−.18	−.12	−.07	1.00					
10. Pay	9.8	6.5	−.03	−.05	.07	−.04	−.21	−.17	−.10	−.14	.21	1.00				
11. Promotion	10.8	7.5	−.15	−.05	−.06	−.10	−.11	−.20	−.09	−.04	.27	.31	1.00			
12. Supervision	44.0	9.1	−.18	−.06	−.05	.13	−.21	.05	−.06	.04	.24	.23	.07	1.00		
13. Co-workers	40.7	10.7	−.07	.05	−.01	−.11	−.13	−.22	−.05	.05	.35	.27	.15	.12	1.00	
Overall job satisfaction																
14. Faces scale	4.0	.9	−.13	−.13	−.18	−.31	−.09	−.11	−.10	−.16	.62	.26	.41	.15	.36	1.00

Note. $r \geq .22$, $p < .01$; $r \geq .16$, $p < .05$. A_{act} = attitude toward the behavior in a given situation; NB_s = social normative beliefs, i.e., perceived expectations of others; BI = behavioral intention; B = overt behavior; JDI = Job Descriptive Index.

It should be noted that scores on the JDI scales were also summed (unweighted) to yield another index of overall job satisfaction. This index correlated −.10 with absenteeism.

The predictive power of combinations of the traditional measures was also examined (Table 2). The multiple correlation of the five JDI scales with absenteeism was .27 (*ns*). Adding the Faces scale to the preceding five predictors increased the multiple correlation to .36 ($p < .05$). This combination of traditional job attitude measures turned out to be the best predictor of absenteeism, accounting for 13% of the variance.

These data suggest that the traditional job attitude measures were relatively more efficacious in predicting absenteeism. Although predictor-criterion correlations of this magnitude are useful to the practitioner, this level of predictive power (i.e., amount of criterion variance accounted for) leaves much to be desired.

An exploratory attempt to increase the predictive power by combining the traditional measures and Fishbein's components yielded a multiple correlation of .37, not significantly better than the combination of just traditional measures (Table 2).

Resignation. BI correlated .39 ($p < .01$) with voluntary resignation (Table 1). A_{act} and NB_s had a multiple correlation of .36 ($p < .01$) with resignation (Table 2). Thus, Fishbein's model accounted for (or predicted) about 13% of the variance in this job withdrawal behavior (Table 2).

Table 1 indicates that of the single traditional job attitude measures, only overall job satisfaction (Faces scale) had a significant correlation ($r = -.16$, $p < .05$) with voluntary resignation. The derived measure of overall job satisfaction (i.e., unweighted sum of the five JDI scales) correlated −.04 with resignation. The correlation ($r = -.14$) between satisfaction with pay (JDI Pay scale) and resignation was not quite significant at the .05 level (Table 1). The rest of the traditional measures had near zero correlations with resignation (Table 1). Thus, the best single traditional measure accounted for approximately 3% of the criterion variance.

Combinations of the traditional measures were only slightly more predictive (Table 2). The multiple correlation of the five JDI scales with resignation was .21 (*ns*). This latter combination accounted for approximately 7% of the criterion variance.

Again, an exploratory attempt was made to increase the predictive power by combining the traditional measures and Fishbein's measures. This resulted in a multiple correlation of .48 ($p < .01$) when BI was included (Table 2). This eclectic combination accounted for 23% of the variance in voluntary resignation.

These data suggest that Fishbein's model was relatively more effective in predicting voluntary resignation. Any conclusions about

TABLE 2

MULTIPLE CORRELATIONS BETWEEN PREDICTORS AND CRITERIA

Measure	Criterion			
	Absenteeism		Resignation	
	R	R^2	R	R^2
A_{act} and NB_s	.12 (.03)	.01 (.00)	.36** (.35)	.13 (.12)
5 JDI scales	.27 (.17)	.07 (.03)	.21 (.06)	.04 (.00)
5 JDI scales and Faces	.36* (.30)	.13 (.09)	.26 (.16)	.07 (.02)
5 JDI scales, Faces, A_{act}, NB_s	.37 (.28)	.14 (.08)	.45** (.38)	.20 (.14)
5 JDI scales, Faces, A_{act}, NB_s, BI	.37 (.27)	.14 (.07)	.48** (.41)	.23 (.17)

Note. One must be careful about the interpretation of the multiple *R*s since they are not cross-validated. For inferential purposes, the values in parentheses have been corrected for "shrinkage" by the Wherry formula. A_{act} = attitude toward the behavior in a given situation; NB_s = social normative beliefs, i.e., perceived expectations of others; BI = behavioral intention; JDI = Job Descriptive Index.

* $p < .05$.

** $p < .01$.

the usefulness and power of either approach, however, must be tempered by a consideration of the amount of criterion variance accounted for.

DISCUSSION

The present study indicated that neither job attitudes nor behavioral intentions were consistently superior in accounting for job withdrawal behaviors. Also, Fishbein's model accounted for less behavioral variance in this field situation than in the prior laboratory studies. Although more field tests are necessary, the present findings reaffirm the view of the prediction of behavior in real organizational settings as a complex situation with many uncontrolled, unaccounted for contributors to criterion variance.

Future attempts to come to grips with this complexity should incorporate more facets of personal and situational information in a comprehensive, but integrated, prediction system. Such an approach is outlined in Newman (1974).

Future research should aim toward further specification of the circumstances under which stronger relationships might be evident. For example, the degree of control the person has over the criterion behavior would appear to be a very important boundary condition. The relationship of attitudes and intentions to behavior is likely constrained to the extent that the person lacks control over the behavior. This may be a partial explanation for the present results in that the study was conducted during a time when the cost of living was high and the labor market tight. Thus, having no better alternatives, many employees had no real choice concerning job withdrawal, perhaps.

Closely related to the problem of lack of control in the field situation and to the degree of control the person has over the behavior of interest, is the fact that employees live in a dynamic world. Thus, a most difficult phenomenon to cope with (let alone control) in predictive, field studies is the change in value of the predictor variables during the time period following their measurement.

If visions of causal relationships dance through our heads when we interpret predictive attitude–behavior correlations, then it is also necessary to entertain the assumption that the attitude (or behavioral intention) remains constant during the intervening time period. Future investigators should test this assumption by closely monitoring the attitude of interest over several points in time prior to obtaining the criterion measure and even at the time of obtaining the criterion measure.

This procedure would also allow examination of the relationship between change (per se) in attitude and the criterion behavior and the relationships of magnitude, rate, and direction of attitudinal change to behavior. These relationships may be stronger and thus more predictive of absenteeism and turnover, for example, than the typical attitude-measured-at-one-time-behavior relationship pursued today.

REFERENCES

Ajzen, I., & Fishbein, M. The prediction of behavioral intentions in a choice situation. *Journal of Experimental Social Psychology*, 1969, **5**, 400–416.

Dulany, D. E. Hypotheses and habits in verbal "operant conditioning." *Journal of Abnormal and Social Psychology*, 1961, **63**, 251–263.

Dulany, D. E. The place of hypotheses and intentions: An analysis of verbal control in verbal conditioning. In C. Eriksen (Ed.), *Behavior and awareness*. Durham, N.C.: Duke University Press, 1962.

Fishbein, M. Attitude and the prediction of behavior. In M. Fishbein (Ed.), *Readings in attitude theory and measurement*. New York: Wiley, 1967.

Fishbein, M. The prediction of behaviors from attitudinal variables. In K. K. Sereno & C. D. Mortensen (Eds.), *Advances in communications research*. New York: Harper & Row, in press.

Kunin, T. The construction of a new type of attitude measure. *Personnel Psychology*, 1955, **8**, 65–78.

Mitchell, T. R., & Biglan, A. Instrumentality theories: Current uses in psychology. *Psychological Bulletin*, 1971, **76**, 432–454.

Newman, J. E. Organized work environments and human behavior: A synthesis of the literature and a framework to guide research and practice. Unpublished manuscript, Bloomington, Ill.: State Farm Mutual Automobile Insurance Company, 1974.

Osgood, C. E., Suci, G. J., & Tannenbaum, R. H. *The measurement of meaning*. Urbana, Ill.: University of Illinois Press, 1957.

Porter, L. W., & Steers, R. M. Organizational, work and personal factors in employee turnover and absenteeism. *Psychological Bulletin*, 1973, **80**, 151–176.

Smith, P. C., Kendall, L. M., & Hulin, C. L. *Measurement of satisfaction in work and retirement*. Chicago: Rand McNally, 1969.

(Received October 15, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 5, 616-622

INDIVIDUAL DIFFERENCES AND REACTIONS TO JOB CHARACTERISTICS[1]

JOHN P. WANOUS[2]

Graduate School of Business Administration, New York University

Three different methods of measuring individual differences were evaluated as moderators of employee reactions to job characteristics. The three methods are urban versus rural background, strong versus weak belief in the Protestant Work Ethic, and high versus low strength for "higher order" needs. Each of these three methods was used as a moderator of the relationships between job characteristics on the one hand, and specific job facet satisfaction, overall job satisfaction, and job behavior on the other hand. Based on the job satisfaction results, higher order need strength is the most useful way to measure this type of individual difference, followed by the Protestant Work Ethic, and lastly by urban versus rural background of the worker. There were no differences among the three individual differences as moderators of the job characteristics and job behavior relationship.

Although certain job characteristics tend to be more satisfying to employees than others, there are wide individual differences in reactions to them. This research is an examination of three different methods for measuring individual differences relevant for employee reactions to job characteristics. The three methods are (*a*) urban versus rural worker background (Hulin & Blood, 1968), (*b*) strong versus weak belief in the Protestant Work Ethic (Blood, 1969), and (*c*) high versus low desire for "higher order need satisfaction" (Hackman & Lawler, 1971). The framework for comparison among the three individual differences concerns how well each moderates the relationship between the *presence* of certain job characteristics and *reactions* to these characteristics.

Job Redesign: Emphasis on Jobs and People

The present viewpoint is that both job characteristics and employee individual differences must be included in a thorough treatment of work humanization. Historically, however, some researchers have tended to emphasize either jobs or people at the expense of the other.

The Herzberg theory (Herzberg, Mausner, & Snyderman, 1959) as applied by Ford (1969) emphasizes job characteristics rather than individual differences. The theory suggests that jobs with opportunities for achievement, recognition, responsibility, advancement, and growth in competence are those that enhance motivation and job satisfaction. Another example of this approach is the Turner and Lawrence (1965) six "requisite task attributes" (i.e., variety, autonomy, required and optional interaction, knowledge and skill required, and responsibility). More recently, Hackman and Lawler (1971) suggested that intrinsically satisfying jobs should (*a*) allow the worker to feel personally responsible for his work (autonomy), (*b*) provide meaningful outcomes (variety and task identity), and (*c*) provide adequate knowledge of results (feedback). Somewhat more general treatments of work humanization (Sheppard & Herrick, 1972; *Work in America*, 1973) have tended to stress a number of factors in the total work environment, for example, pay, supervision, working hours, and so forth.

In contrast to an emphasis on job characteristics, some have concentrated on various types of employee individual differences. These efforts have proceeded along two lines. The first of these has been to identify dissatisfied workers demographically, for example, race, sex, age, or collar color (Sheppard & Herrick, 1972; *Work in America*, 1973). A second approach, the concern of the present study, has been to identify those employees who derive satisfaction from *enriched* jobs in contrast to those who do not.

---

[1] A shorter version of this article was presented at the 1973 annual meeting of the American Psychological Association in Montreal, Canada.

[2] The author wishes to thank J. Richard Hackman, Edward E. Lawler III, Greg Oldham, and Thomas D. Taber for their comments.

Requests for reprints should be addressed to John Wanous, Graduate School of Business Administration, New York University, 90 Trinity Place, New York, New York 10006.

Measuring Individual Differences: Three Approaches

Alienation from middle-class work norms has been suggested as one way to understand why some employees are satisfied with and perform better on enriched jobs (Blood & Hulin, 1967; Hulin & Blood, 1968). This view of individual differences has been operationally measured in two ways: demographically and with a paper-and-pencil measure of work values.

Demographically, Hulin and Blood (1968) used two variables, white collar versus blue collar and urban versus rural work site location. They concluded that the blue collar–urban cell of their 2 × 2 matrix represents those workers who do *not* derive satisfaction from enlarged jobs. Debate has arisen over the best demographic measure (Hulin, 1973; Schuler, 1973; Shepard, 1973). The issue concerns how to define urban/rural differences, that is, work site location or location of socializing experiences while growing up. The present study is not able to resolve this particular issue, because the urban/rural difference was based on personal background during socialization, not work site location (which was exclusively urban).

The second approach to measurement of middle-class work norms employed a short questionnaire designed to measure belief in the Protestant Work Ethic (Blood, 1969). More recently a much longer scale has been developed (Wollack, Goodale, Wijting, & Smith, 1971), but this was not used in the present study.

The third approach measures an individual's preferences or desires for various amounts of certain job characteristics (Hackman & Lawler, 1971). This shares a common questionnaire methodology with the Protestant Work Ethic, but differs conceptually in that specific desires for job characteristics are measured, rather than general views about the value of work.

METHOD

This research was conducted in an eastern telephone company, using a sample of about 80 newly hired female telephone operators who volunteered. These data are part of a larger study (Wanous, 1972).

Each of the three individual differences was operationalized as follows. First, the urban or rural background of an individual was obtained by questionnaire. Subjects were asked to "describe the area where you lived most of the time you were growing up." Thus, location of socialization, not plant location, was used to measure urban/rural differences. This measure was used by Hackman and Lawler (1971) and seems to make sense theoretically (see the last section of this article). Those who stated they grew up in a large (200,000 population) or medium size city (at least 10,000 population) were classified as urban. Those indicating a smaller size city or a rural area were classified as rural. There is no agreed upon definition of what *urban* or *rural* means, so any definition used is at least somewhat arbitrary.

Second, to measure the Protestant Work Ethic only four of eight items from Blood's (1969) scale were used in the present study. The four items measure what Blood calls the "pro Protestant Ethic" (not the "non Protestant Ethic"). The items are (*a*) "Hard work makes you a better person." (*b*) "Wasting time is as bad as wasting money." (*c*) "A good indication of a person's worth is how well he does his job." (*d*) "If all other things are equal, it is better to have a job with a lot of responsibility than one with little responsibility." The internal consistency reliability for this set is .70 when the Spearman-Brown formula is applied to the average item intercorrelation. A median split was used to classify employees as having a strong or weak belief in this ethic.

Third, the measurement of "higher order need strength" was a composite score on eight items measuring the *preferences* for *specific amounts* of various job characteristics. These items concerned: variety, autonomy, task identity, task feedback, management feedback, challenge, meaningfulness, and use of valued skills. These almost exactly parallel those used by Hackman and Lawler (1971). The internal consistency reliability for this set is .78 when the Spearman-Brown formula is applied to the average item intercorrelation. Again, a median split was used to separate employees into high and low "higher order need strength" groups.

In addition to employees' individual differences, the *presence* of certain task characteristics and *reactions* to the characteristics were measured. The degree to which certain task characteristics were present on this job was measured on the four core dimensions of feedback, variety, autonomy, and task identity (Hackman & Lawler, 1971), and several other job characteristics. Each person was asked to describe his job by rating the amount of each job characteristic present on the job.

Three types of employee reactions were obtained: (*a*) satisfaction with a specific job characteristic, (*b*) global job satisfaction, and (*c*) job behavior (i.e., absenteeism and performance).

Satisfaction with specific job characteristics was measured on the Minnesota Satisfaction Questionnaire, (MSQ) Short Form (Weiss, Dawis, England, & Lofquist, 1967). The MSQ is a different scale from that used to measure the presence of job characteristics, and was located in a different part of the questionnaire. This was intentionally done to reduce the amount of common method variance (Campbell & Fiske, 1959) that might spuriously inflate correlations between measures from the same type of scale.

Item	*Item Content*
1. Description	The number of different things done on the job.
Satisfaction	The chance to do different things from time to time.
2. Description	Dealing with co-workers as part of the job.
Satisfaction	Co-workers get along with each other.
3. Description	Dealing with customers as a part of the job.
Satisfaction	The chance to do things for other people.
4. Description	The opportunities to talk informally with other people on the job.
Satisfaction	Co-workers get along with each other.
5. Description	Left on your own to do your work.
Satisfaction	The chance to work alone on the job.
6. Description	Left on your own to do your work.
Satisfaction	The freedom to use my own judgment.
7. Description	Left on your own to do your own work.
Satisfaction	The chance to try my own methods of doing the job.
8. Description	Finding out from the work itself about your own performance.
Satisfaction	I will get a feeling of accomplishment from the job.
9. Description	Finding out from management about your performance.
Satisfaction	I will get praise for doing a good job.
10. Description	The amount of challenge on the job.
Satisfaction	Being able to keep busy all the time.
11. Description	How meaningful the work is.
Satisfaction	I will get a feeling of accomplishment from the job.
12. Description	The use of important skills and abilities at work.
Satisfaction	The chance to do something that makes use of my abilities.

Overall job satisfaction was measured by summing the responses to two global, direct questions: (*a*) Generally speaking, how satisfied are you with your job? (*b*) How characteristic is this statement of you?: Taking everything into account I am very satisfied with my job. The correlation between these two items was $r = .76$.

Job behavior was measured by absenteeism and performance. Job performance was measured by supervisory ratings of an individual's quality and quantity of performance. These two ratings correlated .59 with each other and were combined (linear sum) into a composite measure of job performance. In addition, absenteeism was used as another measure of job behavior. The number of cases of absenteeism in the first three months on the job was used. Occasions absent rather than days absent were used to discount the effects of single long periods of absence.

RESULTS

Table 1 shows the correlations between the *presence* of certain task characteristics and the specific *satisfaction* with each characteristic. If a particular individual differences measure works as a moderator, one would expect to find significantly more positive correlations for high versus low need strength, strong versus weak Protestant Work Ethic, and for rural versus urban background.

The results in Table 1 indicate that the higher order need strength individual differences measure seems to work the best. Of the 12 relationships shown, 8 are significant (one-tailed tests) at the .05 level or better, 3 show differences close to significance (p values between .05 and .10), and 1 is not significant, although it is in the predicted direction. Using the Protestant Work Ethic scale results in fewer significant differences between groups. Of the 12 relationships shown, 4 are significant at the .05 level or better, 1 is close to significance, and 7 are in the predicted direction but not significant. Using the urban and rural backgrounds of these employees as an individual differences moderator shows the fewest differences between groups. Of the 12 relationships shown, only 1 is significant at

TABLE 1

CORRELATIONS BETWEEN JOB CHARACTERISTICS AND JOB SATISFACTION WITH EACH CHARACTERISTIC

Item	Higher order need strength		Protestant Work Ethic		Urban–rural Differences	
	High need strength ($n=35$–39)	Low need strength ($n=35$–39)	Strong Protestant Ethic ($n=35$–39)	Weak Protestant Ethic ($n=35$–39)	Rural background ($n=28$–32)	Urban background ($n=35$–39)
1	.51**	.20	.49	.24	.38*	−.01
2	.42	.25	.39	.17	.52**	.07
3	.29**	−.12	.18*	−.18	.24*	−.13
4	.35*	.00	.27	.13	.36	.08
5	.54**	.18	.48	.23	.38	.33
6	.75***	.30	.67***	.24	.29	.55
7	.49***	−.18	.38**	−.07	−.08	.33
8	.41*	.12	.24	.20	.26	.34
9	.60***	−.02	.39	.37	.24	.16
10	.64***	−.14	.32	.08	−.01	−.24
11	.63**	.23	.56**	.23	.40	.32
12	.68**	.32	.70**	.33	.68*	.44

Note. All tests are one-tailed t tests for the difference between correlations.
* $p < .10$.
** $p < .05$.
*** $p < .01$.

the .05 level, 5 are close to significance, 5 are in the predicted direction, and 3 are the wrong way.

Separate items for each task characteristic were used in the present context rather than a single composite of all items. This was done because the average item intercorrelations are in the low-to-moderate range. Had these intercorrelations been high, it would have indicated that the items were perhaps not referring to discriminable facets of a particular job. In such a case, simply combining all overlapping items into a composite would seem to be indicated. Presently, however, the average item intercorrelation for the task descriptions was $r = .18$, and for the MSQ satisfaction items $r = .25$. Thus, using separate items seems reasonable.

Table 2 shows the correlations between four core task dimensions (variety, autonomy, task identity, and feedback) and overall job satisfaction. The results using overall job satisfaction are not as strong as with specific job satisfaction measures, but the pattern of results is similar to that in Table 1. When using higher order need strength, two of the

TABLE 2

CORRELATIONS BETWEEN JOB CHARACTERISTICS AND GLOBAL JOB SATISFACTION

Type of task characteristic	Higher order need strength		Protestant Work Ethic		Urban–rural background	
	High need strength ($n=35$–39)	Low need strength ($n=35$–39)	Strong Protestant Ethic ($n=35$–39)	Weak Protestant Ethic ($n=35$–39)	Rural background ($n=28$–32)	Urban background ($n=35$–39)
Variety	.50**	.15	.41*	.10	.46	.45
Autonomy	.59***	−.09	.38	.21	.38*	.02
Task identity	.30*	−.07	.33*	−.05	.34	.09
Task feedback	.41*	.10	.24	.22	.34	.17

Note. All tests are one-tailed t tests for the difference between correlations.
* $p < .10$.
** $p < .05$.
*** $p < .01$.

four relationships shown are significant at the .05 level or better, and the other two are close (between .05 and .10) to significance. When using the Protestant Work Ethic measure, none of the four relationships are significant, two are close to it, and two are in the direction predicted but nonsignificant. When using urban/rural differences, none of the four relationships are significant, only one is close to significance, and three are nonsignificant but in the predicted direction.

The individual differences were also used as moderators of the relationship between the four core task characteristics and job behavior. The results were disappointing. There were virtually no differences among the three individual differences for their usefulness as moderators—none of them worked any better than any others. In fact many of the correlations themselves were not significantly different from zero.

Thus far these three individual differences have been compared to each other as to their effectiveness in moderating human reactions to job characteristics. The three individual differences can also be compared by intercorrelating all three measures directly. The inputs to this analysis are the composite total for those eight items measuring higher order need strength, the composite total for the four items from the Protestant Ethic scale, and urban/rural differences expressed as a 4-point scale ($1 =$ large city, $2 =$ medium size city, $3 =$ small city, $4 =$ rural area). The results show that urban/rural background is not closely related to either of the other two, that is, correlating $+.12$ with the Protestant Ethic and $+.17$ with higher order need strength. In contrast, the Protestant Ethic and the higher order need strength correlate $+.41$ ($p < .01$).

DISCUSSION

The present correlational study was designed to compare three different methods of measuring individual differences relevant for future job redesign experiments. The results indicate that the use of higher order need strength is the best of these three individual differences, followed by the Protestant Work Ethic which showed moderate effectiveness as a moderator variable, and last by the urban/rural difference which was generally ineffective in the analyses performed in this study.

In considering the ineffectiveness of urban/rural background differences, at least two points must be kept in mind. First, only one occupational group, telephone operators, was used in the present study. The question is how to classify this group—blue collar or white collar. On the one hand, this occupation is a low paid, entry-level position—similar to a number of blue-collar jobs. On the other hand, this job is typically performed in an office-type situation, not in a factory or plant environment, and thus is somewhat similar to a white-collar clerical position. How one classifies telephone operators is important, because Hulin and Blood's (1968) suggestion was that urban/rural differences are important only for blue-collar jobs not white-collar ones. The fact that this particular job seems to contain elements of both may in itself be one reason why urban/rural differences turned out as they did in this study.

The second point to remember regarding the ineffectiveness of urban/rural differences is that location when being socialized or growing up was used in this study, not location of work site, which was urban. Hackman and Lawler (1971) had some success with this method, but Shepard (1970) did not.

As moderators of the job characteristic–job satisfaction relationship, the three individual differences differed in their usefulness. As moderators of the job characteristic–job behavior relationship, they did not differ. All were generally ineffective. This is quite similar to what Hackman and Lawler (1971, p. 279) found in their analysis using higher order need strength. The only significant differences they found between high- and low-need-strength groups concerned the relationship between task characteristics and specific job facet satisfaction or other attitude variables such as job involvement. One interpretation of this outcome for their study and the present one is that job behavior was measured via different methods (supervisory performance ratings and occasions of absence) than either task characteristics or attitude variables (self-report questionnaire). The generally lower correlations found for job behavior could be due to less common method variance. A second interpretation is that the correlations were lower because the measures of job performance in the present context are less reliable (test–retest) than the satisfaction measures.

Job satisfaction was more reliably measured than job behavior, and shared a self-report method with the measures of job characteristics and the individual differences. This undoubtedly raised the general level of relationships reported in Tables 1 and 2 in comparison to the job behavior data. However, it is much harder to imagine this affecting the differences between correlations as reported in this study. The more reliable job satisfaction measures allow a better test of individual differences. The results for job behavior are most likely artificially low due to less reliable measures of the variables.

It should be recognized that the subjects here were all female and all telephone operators. Thus, generalizations to other groups must be made carefully. Interestingly, the limitation to only one job raised a question for further analysis, that is, how these three individual differences measures worked in relation to average job satisfaction.

A reasonable prediction is that those individuals who have high need strength, or strong belief in the Protestant Work Ethic, or who are from rural backgrounds should have lower average job satisfaction than the corresponding groups. This is because the nature of the operator's job is such that a person with strong desires for an enriched job will likely be dissatisfied as a telephone operator. It is quite plausible to expect those with high strength for higher order needs to have *lower mean* job satisfaction, but still report positive correlations between enriching task characteristics and their job satisfaction. The same could be said for those who strongly believe in the Protestant Work Ethic or who grew up in rural areas.

To test this hypothesis, each of the three individual differences measures was used as a moderator for comparisons of mean global job satisfaction and mean job satisfaction with each of the 20 MSQ items. No significant differences occurred for either the global measure of satisfaction or for any specific facets of it as measured on the MSQ. However, the pattern of differences was highly significant. For example, when high- versus low-need-strength groups were compared, satisfaction was lower on all 20 items of the MSQ for the high-need-strength group. When strong versus weak belief in the Protestant Work Ethic groups were compared, satisfaction was lower for the strong belief group in 17 of 20 items. Using the sign test (Siegel, 1956) reveals that both patterns are significant beyond the .001 level, using a one-tailed test. There were virtually no differences between rural and urban groups. Thus, the results in this study roughly parallel earlier findings and strengthen them.

A Developmental View of Individual Differences

One way to interpret the role of individual differences in the present study is to view the concept of an *individual difference* as the result of a developmental sequence. One of the earliest determinants of an individual's work needs is the environment of his socialization. In the context of a rural or urban white-collar environment, an individual may be more likely to adopt a set of work values similar to what has been called the Protestant Work Ethic or middle-class work values. As a result of such an upbringing, an individual could develop a general value orientation toward work which emphasizes the importance of work in one's total self-esteem and reinforces the belief that work can hold intrinsic satisfaction.

The consequence of such values is that an individual (when asked) is likely to state his own particular desires for job characteristics in terms similar to the higher order need concept. That is, an individual will prefer or desire a job to have large amounts of such characteristics as autonomy, variety, challenge, and meaningfulness. A job will be desired that has task identity and provides the person with feedback. It is likely that such an individual will respond favorably when holding a job with such characteristics.

The rank ordering of individual differences can now be interpreted in a theoretical context. The fact that higher order need strength yielded the clearest results seems reasonable because it is closest to the employee reactions it moderates. Thus, it is also not surprising that the Protestant Work Ethic showed moderate usefulness as a moderator and urban/rural background the least because each is increasingly farther removed from on-the-job attitudes and behavior.

For the future, those interested in applying behavioral science knowledge to the practical problems involved in job redesign should be careful not to overlook differences among the very individuals whose jobs we are trying to change.

REFERENCES

Blood, M. R. Work values and job satisfaction. *Journal of Applied Psychology,* 1969, **53**, 456–459.

Blood, M. R., & Hulin, C. L. Alienation, environmental characteristics, and worker responses. *Journal of Applied Psychology,* 1967, **51**, 284–290.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 1959, **56**, 81–105.

Department of Health, Education, and Welfare. *Work in America,* Cambridge, Mass.: MIT Press, 1973.

Ford, R. N. *Motivation through the work itself.* New York: American Management Association, 1969.

Hackman, J. R., & Lawler, E. E., III. Employee reactions to job characteristics. *Journal of Applied Psychology,* 1971, **55**, 259–286.

Herzberg, F., Mausner, B., & Snyderman, B. *The motivation to work.* New York: Wiley, 1959.

Hulin, C. L. Reply to Schuler. *Industrial and Labor Relations Review,* 1973, **26**, 853–855.

Hulin, C. L., & Blood, M. R. Job enlargement, individual differences, and worker responses. *Psychological Bulletin,* 1968, **69**, 41–55.

Schuler, R. Worker background and job satisfaction: Comment. *Industrial and Labor Relations Review,* 1973, **26**, 851–853.

Shepard, J. M. Functional specialization, alienation, and job satisfaction. *Industrial and Labor Relations Review,* 1970, **23**, 207–219.

Shepard, J. M. Reply to Schuler. *Industrial and Labor Relations Review,* 1973, 26, 856–859.

Sheppard, H. L., & Herrick, N. Q. *Where have all the robots gone?* New York: Free Press, 1972.

Siegel, S. *Nonparametric statistics.* New York: McGraw-Hill, 1956.

Turner, A. N., & Lawrence, P. R. *Industrial jobs and the worker.* Boston: Harvard University Graduate School of Business Administration, 1965.

Wanous, J. P. An experimental test of job attraction theory in an organizational setting. Unpublished doctoral dissertation, Yale University, 1972.

Weiss, D. J., Dawis, R. V., England, G. W., & Lofquist, L. H. *Manual for the Minnesota Satisfaction Questionnaire.* Minneapolis, Minn.: University of Minnesota Industrial Relations Center, 1967.

Wollack, S., Goodale, J. G., & Wijting, J. Development of the survey of work values. *Journal of Applied Psychology,* 1971, 55, 331–338.

(Received September 24, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 5, 623–632

SOME PERSONALITY CORRELATES OF THE RELATIONSHIPS BETWEEN INDIVIDUALS AND ORGANIZATIONS

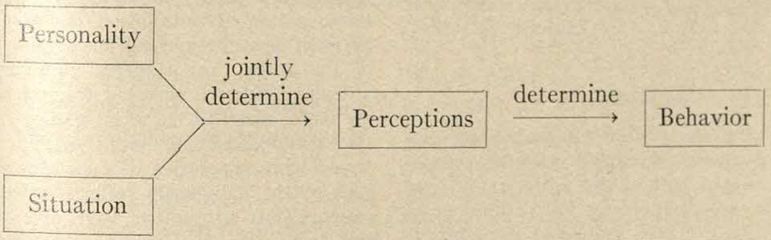
H. RUSSELL JOHNSTON, JR.[1]

Babcock Graduate School of Management, Wake Forest University

Using the Thematic Apperception Test and nondirective interviews with 39 professionals employed by a consulting firm, the study investigates the link between three personality dimensions and the individual's perception of his/her relationship with the organization. Results show that individuals with active or high task orientation perceive a more positive relationship with the organization generally and with specific features of the organization than do passive or low task peers. Differences along the interpersonal dimension were not significant. The study raises questions concerning more effective matching of individuals and organizations.

Much of the current theory about organizational functioning alludes to, but deals only summarily with, the effects of different modalities of individual personality orientation on behavior in organizations. Specifically, there has been little attempt to determine whether particular attributes of the individual's personality have significant effects on the nature of the relationship of the individual to purposive organizations and therefore on his or her behavior in organizational settings. Sharing the conventional assumption that basic personality orientation influences the relationship between the individual and the organization, the present study investigated the nature and strength of these influences.

The model upon which this study is based is that basic aspects of the individual's personality importantly affect the way that the individual perceives his/her environmental situation, and that perception in turn influences behavior.



The research reported in this article is attempting to test the hypothesis that individuals faced with a common environment perceive it differently, and that these differences in perception are strongly influenced by particular dimensions of the individual's personality.

Three personality dimensions were chosen for study. These were activity–passivity, task orientation, and interpersonal orientation.

Activity–passivity. This dimension relates directly to the capacity of the individual to establish and maintain a satisfying and productive relationship with his environment. Robert White (1963) has stressed the importance of the individual's capacity to take effective action (as opposed to feeling acted upon) in achieving a sense of competence or feelings of effectance. White develops the thesis that the need to take effective action to achieve a sense of efficacy is an independent ego drive similar in strength and importance to the sexual and aggressive id drives. If White's theory is correct, one would expect that the individual's capacity for effective and sustained action would have an important influence on the quality of his relationship to purposive organizations.

---

[1] Requests for reprints should be sent to Russ Johnston, Babcock Graduate School of Management, Wake Forest University, Winston-Salem, North Carolina, 27109.

Task and interpersonal orientation. These two dimensions are concerned with the means by which the individual attempts to establish and maintain relationships with his environment. The person with strong task orientation seeks to establish his sense of competence through manipulation of the object environment, while the person with a high interpersonal orientation relies to a greater degree on interpersonal relationships as a means of achieving satisfaction of his needs for effectance. Thus these two personality dimensions deal with approach modalities, whereas activity–passivity deals with approach capacity.

The present study attempts to determine whether variations among individuals along these three personality dimensions relate to variations in perception of the quality of the relationship between the individual and the organization (I/O relationship). Thus measures of the strength of the individual's activity and task and interpersonal orientation become the independent variables, and measures in perception of the quality of the I/O relationship are taken as dependent variables.

METHOD

Subjects for the study were all professional employees of a five-year-old, rapidly growing consulting company. About half of the subjects (19) had been with the organization for over three years, while 20 subjects had been employed for terms of six months to two years.

To develop a measure of the independent variables (scores on personality dimensions), the Thematic Apperception Test (TAT) was used. In selecting this test, the author was sharing the assumption/conclusion of others who have used it extensively (e.g., Henry, 1956; McClelland, 1958; Murray, 1938) that the test enabled one to identify basic and enduring aspects of a subject's personality as revealed in the projective fantasy responses to the cards. Henry (1956) described the TAT stories as portraying the subject's "usual definition of and usual manner of responding to (such) reality events [p. 10]." Six cards (numbers 1, 2, 7BM, 13B, 14, and 17BM) were selected from the set developed by Murray (1943). Selection was based on an attempt to balance the stimulus properties of the cards among the three personality dimensions of interest. Subjects were given the standard instructions to tell an imaginative story based on the picture on the card. All stories were recorded on tape for later analysis.

A scoring manual was developed to provide guidelines for counting and classifying imagery produced in the stories which related to the three personality dimensions of interest. Each story protocol was classified (scored) on a scale of 1 to 10 based on the nature of activity–passivity imagery produced. The average of the scores on the six stories was used as the measure of the subject's proclivity toward activity or passivity. A high score represented a relatively active orientation. Scores on the task and interpersonal dimensions were based on the presence and elaboration of task and interpersonal imagery. Each story was scored separately. The sum of the task and interpersonal scores on the six stories was used as the measure of task and interpersonal orientation.

TAT protocols were scored independently by two trained scorers. Reliability correlations were .90 for the task dimension, .84 for interpersonal, and .94 for the activity–passivity dimension.

Data on the quality of the I/O relationship were gathered by conducting a relatively nondirective interview with each of the 39 subjects. At the beginning of the interview the subject was told that the study involved the relationship of the individual to an organization and was asked to describe his own relationship to the host organization. He was requested to particularly discuss any changes that had occurred in the I/O relationship, their causes and effects, and the adaptions which the changes had made necessary or desirable. At this point the sophistication of the subjects and the milieu within the host organization was important to the success of the study. The subjects were generally highly articulate and well educated (most had done graduate work, many in the social sciences). Furthermore, the atmosphere and norms within the company were highly supportive of open discussion of issues such as the individual's feelings toward the organization. Thus all subjects were able and willing to talk freely for a period of one-half hour to an hour and a half without substantive intervention by the interviewer. These interviews were tape-recorded.

The protocols of the unstructured interview were subjected to a two-pass content analysis. The first pass was for the purpose of obtaining an exhaustive list of all the types of comments made by the subjects about the organization and their relationship to it. This procedure resulted in a list of 264 statements which were then grouped into 25 comment categories. (A comment category included all statements by the subject about the effect of a particular feature of the organization on his/her I/O relationship.)

A four-page scoring sheet was developed which included all comment types grouped within categories along with provisions for indicating positive and negative affect associated with a statement. The scoring system also indicated whether a subject used task, interpersonal, or engage–withdraw imagery in discussing a feature of the organization or his I/O relationship. On any given statement a subject might be scored for positive and/or negative affect about a feature of the organization and/or its effect on the I/O relationship. Descriptive statements made without evaluative comments did not receive an affect score, although they might have been scored for presence of imagery.

A second pass of the tape-recorded protocols was then made to score each interview. At the completion of this pass, the net number of positive minus negative comments made by the subject in each comment category was used as a measure of the subject's attitude toward that aspect of the organization or I/O relationship to which the comment category related. In addition, the number of occurrences of task, interpersonal, active, or passive imagery was accumulated for each imagery class.

RESULTS

The scoring of TAT protocols resulted in the data (Table 1) for the three personality dimensions. The high intercorrelations between the active-passive dimension and both the task and interpersonal dimensions were expected since the former dimension is conceived as a measure of the individual's capacity for acting upon his environment, while the two latter dimensions are viewed as channels or means for interacting or acting upon reality. A check for relationship between the scores on the three dimensions and demographic factors such as age, company tenure, sex, marital status, job level, and educational level yielded no relationships even approaching significance.

Nonparametric correlations were calculated to test the relationship between personality dimension scores and the total number of occurrences of active, passive, task, and interpersonal imagery, and positive and negative comments within the unstructured interview. Table 2 shows these results.

TABLE 2

RELATIONSHIP BETWEEN THEMATIC APPERCEPTION TEST AND INTERVIEW IMAGERY AND COMMENTS

Interview data	Thematic Apperception Test dimensions		
	Active-passive	Task	Interpersonal
Imagery			
Engage	.73**		
Withdraw	−.53**		
Task		.65**	
Interpersonal			.19
Comments			
No. positive	.46**	.39*	.31
No. negative	−.22	−.26	−.11
Net (positive − negative)	.45*	−.42**	.20

Note. Spearman nonparametric test in the Statistical Package for the Social Sciences (SPSS) was used.
* $p < .025$ (one-tailed test).
** $p < .01$ (one-tailed test).

These results suggest that imagery used by the subjects in their fantasy responses to the TAT cards is rather strongly related (in the active–passive and task classes) to the imagery with which they perceive and react to their organizational environment and their own relationship to it. Such a relationship is, of course, the raison d'être for the TAT, but, frankly, the investigator was startled by the strength of the relationships on activity and task dimensions. These results lead one to suspect that the weaker relationship on the interpersonal dimension is attributable to problems in that section of the scoring man-

TABLE 1

THEMATIC APPERCEPTION TEST SCORES

Personality dimensions	Range	*M*	*SD*	Intercorrelations		
				Activity–passivity	Task	Inter-personal
Activity–passivity	1.2–8.5	4.2	2.1	—		.31
Task orientation	1.0–22	9.4	6.0	.48	—	
Interpersonal orientation	0.0–17	6.2	4.1		.20	—

ual or to the weak interpersonal stimuli in the TAT cards used in the study.

The results in Table 2 also suggest that at the aggregate level, those subjects with strong activity and/or task orientations feel more positively about the host organization and their relationship to it than do their peers with a more passive and/or lesser task orientation.

Subjects were then divided into two groups on each of the three dimensions. Those with a score of 5.0 or greater on the activity-passivity dimension were classified as Active; lower than 5.0, Passive. Similarly task and interpersonal scores of 10 or higher resulted in a subject being classified as high on that dimension. This procedure resulted in the following breakdown:

N(active) = 18
N(high task) = 19
N(high interpersonal) = 9

N(passive) = 21
N(low task) = 20
N(low interpersonal) = 30

The disparity in Ns on the interpersonal dimension is, of course, statistically unfortunate, but there was a natural break in the scores from 10 to 8 which, upon investigation, seemed to provide the best line of demarcation between high and low interpersonal TAT protocols.

In order to gain a finer grain picture of the nature of the I/O relationship for members of high and low scoring groups on the TAT dimension, each of the 25 comment categories was examined. The dependent variable used was the average number of positive minus negative comments made in each category by members of the group, for example, $\sum_{i=1}^{N}$ (numbers of positive comments about "participation in decision making" minus the number of negative comments about participation)/N; where N equals the number of subjects in the TAT scoring group. This statistic is referred to as the mean positive balance of comments. Results for the 25 comment categories with a breakdown by personality grouping are shown in Table 3. (No attempt was made to test for significance of difference between the means for high and low interpersonal scorers because of the small sample size of high scores and the disparity between the two Ns.)

The data in Table 3 tend to support the general conclusion from the aggregate results presented in Table 2 in that those subjects with relatively active or high task or interpersonal orientations showed a more positive perception of specific features of the organization and their relation to it than did their more passive, or low task, or interpersonal peers. Active subjects were more positive on 19 of the 25 categories (eight of the differences between the means were found to be significant at the .025 level or higher); high task subjects were more positive on 20 of the 25 categories (eight were found significant at the .05 level or higher); and high interpersonal subjects were higher on 18 of 25 categories (none significant). The nature of these perceptual differences is explored in following sections.

Organization Climate

Active versus passive subjects. Active subjects saw the organization as highly unstructured, open, and egalitarian. They tended to experience substantial support for innovative ideas, recounted no unreasonable difficulty in pursuing projects in which they became interested, and found the atmosphere professional and intellectually stimulating.

Passive subjects made a number of comments about the pervasiveness of office politics. They perceived a wide gulf between managers and other professionals which prevented a free flow of ideas, particularly the flow up the hierarchy. A number of passive subjects objected to the chaos in the organization and complained that they were unable to work in such an atmosphere.

High task versus low task subjects. Subjects who scored high on task orientation expressed substantial satisfaction with the de-

TABLE 3

MEAN POSITIVE COMMENT BALANCE BY COMMENT CATEGORY AND PERSONALITY GROUPING

Comment category	Personality grouping					
	Active $N = 18$	Passive $N = 21$	Task		Interpersonal	
			High $N = 19$	Low $N = 20$	High $N = 9$	Low $N = 30$
Commitment to organization	3.7	4.7	5.7	2.9	5.1	4.0
Organizational climate	3.3***	−2.2	3.0**	−2.2	3.3	−0.5
Evaluation of own performance	3.3**	0.2	2.8	0.6	3.2	1.7
Personal goals and strategies	3.8***	−0.3	2.3	0.6	3.6	1.0
Organizational structure, systems, practices	−2.9	−4.1	−3.6	−3.5	−2.6	−3.9
Organizational goals and incentives	1.7	−0.4	1.9**	−0.8	−0.4	0.8
Quality of organization's product	−1.2	−2.4	−1.7	−1.9	−3.3	−1.4
Top management style	−3.0	−1.7	−1.9	−2.7	−6.1	−1.2
Superior-subordinate relations	−0.6**	−5.2	−2.0	−4.2	−2.8	−3.2
Organizational role expectations	1.8***	−1.0	1.4**	−0.8	0.9	0.1
Individual's role expectations	2.0	−0.4	1.7	0.3	1.2	0.5
Organizational changes	2.5	−0.9	3.2**	−1.7	3.1	0.0
Support for initiative	6.2***	1.1	4.6	2.3	5.9	2.7
Reward systems	0.2***	−2.6	−0.1**	−2.6	−1.1	−1.4
Participation in decision-making	−0.7	−0.2	−0.8	−0.1	0.1	−0.6
Pressure on the job	−0.3**	−2.7	−0.6*	−2.6	−1.6	−1.6
Social atmosphere	0.1	0.1	0.4	0.2	0.3	0.0
Sales orientation	−0.2	−0.4	−0.1	−0.5	−0.2	0.3
Congruence of individual and organization expectations	−0.2	−1.9	0.0*	−2.1	−1.1	−1.1
Formal system changes	0.2	−0.1	0.1	0.0	0.1	0.0
Climate changes	0.1	−0.1	−0.2	0.1	0.2	−0.1
Dependence on superior	0.1	1.1	0.3	0.9	0.9	0.5
Bad faith in I/O relationship	−0.3	−1.1	−0.6	−0.9	−0.7	−0.8
Commitment to nature of work performed	1.2	0.9	1.3	0.8	0.9	1.1
Desire for promotion	0.3	−0.3	0.4	−0.5	0.1	−0.1

Note. Tests for significance between the means used the t statistic for small sample sizes with appropriate reduction in degrees of freedom for unequal Ns. A one-tailed test was used.

* $p < .05$.

** $p < .025$.

*** $p < .01$.

gree to which the organization's climate facilitated accomplishment of their project goals by permitting access to the necessary resources. Low task scorers generally expressed criticisms of the climate in much the same manner as passive subjects.

Evaluation of Own Performance

Active versus passive subjects. Active subjects generally felt reasonably satisfied with the quality of their work, with their personal professional development since joining the company, and with the recognition and rewards which they perceived as related to their work. Many active subjects commented that their work had had substantial impact upon the company and/or its clients. Perhaps most importantly, many reported instances in which they had altered the work situation to facilitate achievement of their objectives. Passive subjects were not only less satisfied with their own performance, but frequently expressed strong resentment toward perceived management blunders which had prevented them from achieving success, resulting in frustration and resentment. They also expressed far more concern over how others evaluated their work than did active subjects.

Personal Goals and Strategies

Active versus passive subjects. This category proved particularly valuable in understanding the differences between active and passive subjects. Subjects near the extreme activity range reported (almost unanimously) a clear and well-articulated set of personal goals and realistic strategies for achieving these goals. Some envisioned a long association with their current employer; others saw their current job as a temporary relationship during which they planned to achieve specific advantages and progressions. The more passive subjects, on the other hand, were not only imprecise or unclear in goal definitions, but many stated that their lack of goal direction was intentional and advantageous to them.

Organizational Goals and Incentives and Organizational Changes

High task versus low-task subjects. These two comment categories are closely related in that the rapid growth of the host organization had resulted in a series of related changes in the organization's structure and internal systems. The most prominent of these changes involved attempts by management to gain sufficient control over the organization to insure an orderly and reasonably efficient pursuit of the organization's goals as conceived (primarily) by the senior members of the organization. The changes (by their nature and effects) had two important results. First, they tended to clarify, objectify, and make more internally consistent the company's goals. Second, they tended to introduce order into the project work situation (task). The increased clarity of goals and the more systematic and methodical work procedures were greeted with enthusiasm by the high task scorers who reported that they facilitated improved task performance. Low task scorers tended to ignore or denegrate the facilitative aspect of the changes and to characterize them as "bureaucratic red tape." Low task subjects also perceived that organization goals were too narrowly conceived and related almost exclusively to profit maximization while ignoring such goals as personal development and socially oriented objectives.

Superior-Subordinate Relationships

Active versus passive subjects. While all subject groups tended toward a negative balance of comments about the character of superior-subordinate relationships, active subjects were markedly different in one important respect. They shared a pervasive feeling that middle managers in the company often made mistakes or were poorly prepared to do managerial jobs, but many active subjects disclosed that they had found or developed ways of compensating for or working around these problems in order to achieve their own goals. Few passive subjects even alluded to constructive responses.

Organizational Role Expectations

Both active and high task subjects expressed feelings that the organization expected a high degree of commitment, initiative and innovativeness, and adherence to professional standards of quality in performance. They saw these expectations as reasonable and regarded both the extrinsic and intrinsic rewards available within the organizational environment as being commensurate with the company's expectations. Passive and low task scorers shared the same recognition of the manifest demands of the company as the active and high task subjects. They were more likely, however, to regard these expectations as either phony ("They say they're looking for initiative and innovation, but they really aren't") or unwarranted ("The demands here are excessive").

Support for Initiative

All subject groups showed a positive comment balance on the category related to the organization's support for initiative. There was, however, a very significant difference between active and passive subjects. All active subjects shared the view that the preponderance of informal attitudes and formal procedures were generally favorable to the individual exercise of initiative. Most active subjects saw the situation as actively calling for, even demanding, initiative. Active subjects also frequently reported incidences in which they had exercised initiative and experienced favorable responses.

The passive subjects were mixed in their responses within the category. Some expressed recognition of support for initiative but reported that they had not felt a need or wish to take initiative. Others expressed a variety of feelings that the apparent support of initiative was a ruse, either intentional or unintentional. A few were quite bitter in denouncing the manifest support as a willful fraud and recounted the frustrating experiences of themselves and others to support the charge.

Pressure on the Job

In this category there was again agreement among members of all subject groups about the objective existence of a relatively high level of pressure in the work environment. However, the remarks of the active subjects cited much of the pressure as internally generated and at times functional. High task scorers agreed to some extent with the functional assessment, but saw the pressure arising from the nature of the work being performed (project work with stringent deadlines under conditions of high uncertainty and a lack of control over important factors). Passive subjects, on the other hand, attributed the pressure to management failures and a tendency to manipulate the junior professionals through inducing guilt.

The Emerging Picture

A pattern of responses emerges from the data which takes the following form:

Active subjects. Subjects with a relatively active personality orientation perceive fewer and less severe problems and constraints within the organizational environment than their passive co-workers. Perhaps more importantly, they regard these problems and constraints as matters which called for adaptability and increased effort in order to deal successfully with them. Many reported, along with their recitation of the existence of problems, proactive attempts at solution. Though not all reported solution attempts were described as successful, failures were generally discussed with some equanimity and often with intent to try again or continue trying. Equally notable to their relative persistence was the clarity, preciseness, and realistic nature of the goals and goal strategies expressed by the active subjects.

Passive subjects. Passive subjects displayed quite clearly the familiar constellation of passive–aggressive orientation. Not only did they see many of the problems as more severe than did their more active counterparts but in dwelling on the problems, they expressed far more bitterness and anger. Senior members of the firm (authority figures) were cited as the usual cause of problems and there was a greater tendency to impute capricious and manipulative motives to managerial acts. Goals, if mentioned, were less immediate and less reality oriented. The dissatisfaction of these subjects extended to their own performance as well as to the organizational situation.

High task scorers. These subjects were positive and explicit about aspects of the environment which facilitated achievement of work objectives. They made more comments about specific projects, and their criticisms were often tied to problems related to their work within the company.

Low task scorers. Low task scorers were not only more negative toward the organization and their relationship to it than were high task subjects, but also their criticisms were more diffuse and general in nature. They expressed a more pervasive and elusive sense of uneasiness and frustration with the climate, goals, systems and practices, superior–subordinate relationships, expectations, etc. They cited fewer specific incidents which caused dissatisfaction and spent less time discussing their own project assignments. In perception terms, they discussed the ground rather than the figure.

CONCLUSIONS

First, it is necessary to look again at the nature of each of the two sets of data. The TATs were administered prior to disclosure of the subject of the study in order to produce a set of data which revealed characteristic ways in which the individual's personality defined and dealt with reality. The attempt was to minimize the influence of stimuli from the organizational situation on this set of data.

The second set of data from the interview reflects the introduction of organizational stimuli by the investigator. However, the unstructured nature of the interview was an attempt to prevent the investigator's own perceptions or biases concerning the objective reality of the setting from tainting the interview data. Thus this set of data represents the subject's perceptual field of his/her particular organizational setting as determined by the interaction of stimuli from both individual personality and "organizational reality" (as per the model early in this article).

While the data presented do suffer from the imprecision of classification and the inability to isolate causation commonly found in exploratory field studies, the results do suggest that the personality dimensions studied (particularly the active/passive and task dimensions) as measured by the TAT bear a strong relationship to the perceptual field data gathered during the interview. Put another way, the personality variables appear to account for (in the regression sense) some of the variation in affective perception of the organization and the I/O relationship. If, as phenomenological psychologists suggest, the perceptual field determines behavior, then the logical extension of this finding is that these aspects of personality would also account for variation in behavior in organizational settings. Research currently underway is attempting to test, in a variety of settings, the validity of this extension. Given the findings of the present study, it is now possible to examine these relationships in a more directed and precise manner.

The present results seem to fit the White (1963) "sense of competence" model quite neatly. The consistent reports by active subjects that they were able to have an appreciable and positive (at least from their point of view) effect on the environment within the organization appear to be quite closely related to the greater satisfactions which they express with the organization and their own situation within it. There is qualitative and quantitative support for the assertion that these relatively active people are deriving important satisfactions from a sense of efficacy. As a consequence they had a higher regard not only for their own performance but for the organization which provided a setting for—perhaps even facilitated—their successful maintenance of feelings of efficacy.

Passive subjects expressed feelings of being acted on by the managers and other forces in the organizational milieu. These feelings are the antithesis of feelings of efficacy and lead directly to the passive–aggressive attitudes against the organization and phenomena over which these subjects felt no capability for control.

As was mentioned earlier, the task and interpersonal dimensions have to do with the means by which the individual relates to his environment in search of satisfaction (whereas activity–passivity corresponds to the strength and direction of the drive). High task scorers regularly related their goals to task performance and responded favorably to those organizational mechanisms which tend to reduce ambiguity and uncertainty surrounding their tasks. They responded favorably to assignments within what Thompson (1967) calls the "technological core" of the organization and sought buffering from both external and internal sources of uncertainty.

Existence of Multiple Climates

Perhaps one of the most intriguing—though admittedly serendipitous—findings was that two rather distinct organizational climates existed within the small organization where the study was conducted. One climate was primarily the realm of those employees who had joined the organization in its early, formative period. This group, despite being distributed among all hierarchical levels within the firm, had maintained an organic type of climate which operated primarily through reliance on informal systems. The newer employees who had joined the organization at a time when its size had forced top management to install and rely on more formal organizational mechanisms for control, encountered a more rigid, formal system.

The discovery of two rather distinct climates within the organization made it possible to investigate the relationship between the personality dimensions and I/O perception under varying organizational or situational conditions. For this purpose, the 19 subjects who had been with the company for over three years were classified as "first generation employees." The 20 subjects who had been employed for periods ranging from six months to almost two years were grouped as "second generation." A check revealed that the distributions of TAT scores were quite similar in mean, range, and variance for the two generational groups. Thus the personality measures (independent variables) were considered to be unrelated to the generational breakdown. While generational differences in perception may be attributable to shifts in selection criteria, both personality tests and subjective interview information suggest that conscious company efforts at maintaining consistency in personnel selection had been reasonably successful.

Relation of Climates to I/O Perception

A look at the independent variable (perception of I/O relationship) revealed an interesting pattern. On the one hand, the differences in I/O perception between active and/or high task and passive and/or low task, second-generation subjects were quite similar in direction and magnitude to comparable first-generation groups. For instance, both active first-generation and active second-generation subjects displayed more positive perceptions than did their more passive generational peers, and they did so to similar degrees. Inference—the relationship between the personality dimensions and I/O perception appears to hold over at least some range of variation in situational variables.

On the other hand, first generation active and/or high task subjects expressed consistently more positive perceptions than second-generation active/high task employees (first-generation active employees were more positive on 21 of 25 comment categories). Similarly first-generation passive/low task subjects were more positively disposed than were second-generation passive/low task employees on 14 of 25 categories. Second inference—organizational climate variables also relate to the dependent variable (perception of the I/O relationship).

These findings lead one toward the tentative formulation that, at least over some range of variation in organizational mechanisms (stimuli), the personality dimensions studied bear a steady relationship to perception of the I/O relationship. However, I/O perception does vary with variations in organization climate. I/O perception is indeed found to vary as a joint function of both the personality variables studied and organizational climate variables.

A number of questions still remain in the intriguing and interesting category:

1. Would varying the organizational setting over a wider range of variation narrow or even reverse the results? Are there organizations in which passive subjects would express more positive perceptions of the I/O relationship than would their more active peers? Intuitively it would seem that the answers would be "yes," at least in the extremes of organizational inflexibility. However, the range of variation in organization climate in this study did not produce such a result. While increased rigidity related to less positive attitudes for active subjects, they remained more positive than for passive peers under the same conditions.

2. To what degree do differences in personality dimensions among subjects lead them to create the situations which they perceive (e.g., Does the passive individual behave in his/her interactions with others in the environment in such a way that he/she creates an environment, the reactions of others, that is objectively different in important ways than that faced by more active peers?) or are the differences reported in this study entirely a matter of perception? One of the current criticisms by Argyris (1972) of organizational sociology is that man is treated as "a mechanistic closed-system. Man is not conceptualized as proactive but as passive with little influence on the organization [p. 72]." As the investigator listened to active subjects describe their I/O relationship, his consistent impression was that they were de-

scribing processes in which they altered and shaped their objective environment in a proactive manner. The converse also appeared to be true. Though the processes are far more subtle, it appeared that passive employees, through their behavior, did create a more controlling, unyielding environment. A more precisely calibrated study is necessary to determine whether these effects of the individual personality on its environment do exist and if so what forms they take.

3. If there are types of organizational climates which are more suited to particular types of individuals, how can social science methods aid in better matching of individuals to organizations?

4. Are there positive means to affect or alter the activity–passivity, task, interpersonal, or other dimensions of individuals to aid in their adjustment to purposive organizations? Or alternatively, what dimensions of the organization can the manager adjust in order to accommodate the differences in personality dimensions among employees?

Each of these questions poses needs for a great deal of thoughtful and well-executed study before convincing answers can be posed. However, both our society's quest for more meaning and satisfaction from work and the angst experienced by the passive person in an organization demanding proactive behavior call for answers.

REFERENCES

Argyris, C. *The applicability of organizational sociology*. Cambridge, Mass.: University Press, 1972.

Henry, W. E. *The analysis of fantasy*. New York: Wiley, 1956.

McClelland, D. C. The importance of early learning in the formation of motives. In J. W. Atkinson (Ed.), *Motives in fantasy, action, and society*. Princeton, N.J.: Van Nostrand, 1958.

Murray, H. A. *Explorations in personality*. New York: Oxford University Press, 1938.

Murray, H. A. *Thematic apperception test*. Cambridge, Mass.: Harvard University, 1943.

Thompson, J. *Organizations in action*. New York: McGraw-Hill, 1967.

White, R. *Ego and reality in psychoanalytic theory*. New York: International Universities Press, 1963.

(Received July 16, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 5, 633-637

COMPARISON OF JOB EVALUATION METHODS: A "POLICY-CAPTURING" APPROACH USING THE POSITION ANALYSIS QUESTIONNAIRE[1]

DAVID D. ROBINSON[2]
Boise, Idaho

OWEN W. WAHLSTROM
Utah Power and Light Company

ROBERT C. MECHAM[3]
Department of Business Administration, Utah State University

A policy-capturing method of job evaluation, which involved the multiple regression weighting of numerical job analysis data obtained using the Position Analysis Questionnaire (PAQ), was used to predict compensation rates in two studies. In Study 1, compensation rates were derived and compared for 19 municipal government jobs using the policy capturing and four conventional methods. The rates derived using the five methods intercorrelated between .82 and .95. In Study 2, PAQ data and job evaluation points previously determined using a point method were obtained for 131 jobs in a utility company. A stepwise regression analysis, used to select and weight PAQ factor scores to predict the job evaluation points, yielded a multiple correlation of .90. Because of comparable results and lower cost, the policy-capturing method was adopted in both organizations.

The use of a "policy-capturing" approach to discover the relationship between judges' decisions and the quantifiable data used to arrive at those decisions has been used in a variety of contexts (e.g., Doherty & Keeley, 1972; Goldberg, 1965, 1968; Slovic & Lichtenstein, 1971). The policy-capturing method typically involves a multiple regression analysis in which data on variables available to judges are considered as predictors, with the judges' decisions as criteria. Several investigators (e.g., Champagne & McCormick, 1964; Finn & Lee, 1972; Madden, 1963; McCormick, Jeanneret, & Mecham, 1972) have applied this approach to a range of job evaluation problems in which job, incumbent, or judge characteristics expressed in quantitative form as predictors are analyzed to statistically capture the policy which led to rates of compensation established by some job evaluation process. The predictor weights, in equation form, resulting from a regression analysis presumably reflect this policy and could be used thereafter to predict compensation rates if predictor data could be obtained.

Following the development of the Position Analysis Questionnaire (PAQ)[4] (McCormick, Jeanneret, & Mecham, 1969), a structured job analysis instrument, data about the characteristics of a wide variety of jobs were obtained. Multiple correlations between various combinations of job characteristics and wage and salary rates ranged between .82 and .90 for a sample of 340 jobs (McCormick et al., 1972; Mecham & McCormick, 1969). Later, in a cross-validation study, L. R. Taylor (personal communication, 1970) obtained a correlation of .93 between existing salary rates and predicted rates for 79 insurance company jobs analyzed with the PAQ; and Harris and McCormick (1973) successfully applied a policy capturing approach in the analysis of rates of compensation for U.S. Navy jobs.

The present report extends the use of the policy-capturing approach to an operational setting involving a medium-sized municipal government (Study 1) and a utility company (Study 2), and compares the results obtained with those of several traditional methods of job evaluation.

---

[1] The authors express appreciation to Jared Volmar, Glen Rawlins, and the students involved in the data collection and analysis and to the two organizations represented for funding the research.

[2] A consulting psychologist in private practice.

[3] Requests for reprints should be sent to Robert C. Mecham, Department of Business Administration, Utah State University, Logan, Utah 84322.

[4] The Position Analysis Questionnaire is available from the University Bookstore, West Lafayette, Indiana 47906.

STUDY 1

Method

Five methods of job evaluation were employed in deriving compensation rates for 19 benchmark jobs from the city of Boise, Idaho (population 74,990). The benchmark jobs were selected on the basis of representation throughout the compensation range, clarity of function, and likely occurrence in other municipal governments.

The methods of evaluation included the following: (*a*) a ranking method in which the jobs were ordered in terms of overall worth, (*b*) a factor comparison method in which rates of a sample of equitably compensated jobs were divided among compensable factors (e.g., knowledge required, skill, hazards, etc.) which resulted in a reference scale against which other jobs were compared and their rates established, (*c*) a point method in which the job characteristics were matched with various degrees of compensable factors having designated point values from which total points and compensation rates were determined, (*d*) a policy capturing method which predicted rates by use of a regression equation to weight PAQ data, and (*e*) a wage and salary survey of similar municipalities to determine the median compensation rate for each job. The survey data were viewed as a criterion against which the results of the other methods could be compared especially if the methods yielded dissimilar results.

As a first step in the evaluation procedures, job descriptions were prepared or revised and coded using the *Dictionary of Occupational Titles* (United States Department of Labor, 1965, 1967).

Secondly, mayors of all cities in the 48 contiguous states having populations within ± 10,000 of that of Boise and which were located 50 miles or more from a large metropolitan area were asked to participate in the wage and salary survey. The survey asked for current compensation rates for 17 jobs and area wage survey data (United States Department of Labor, 1970) were available on two additional benchmark jobs (account clerk and secretary) which were adjusted to the current compensation level by adding an 8% cost of living factor. Of the 31 cities contacted, 21 returned the requested data.

As a third step, job analyses of almost all of the city's 168 full-time jobs were performed using the PAQ by one or two incumbents, a supervisor, and one or two upper-division college students who worked with both incumbents and supervisors in the field. These analyses involved the rating of the degree to which each of the 194 job elements in the PAQ was part of the job. Interpretational assistance of job element meaning was provided at times but not to the extent that independence of observation was violated. The PAQ data for each job were then averaged and job dimension (factor) scores were calculated from the job element scores using a computer program located at Utah State University. Wage and salary estimates were then made by weighting the job dimension scores using a general nonlinear regression equation previously derived from a sample of 531 jobs from approximately 50 different organizations (McCormick, Mecham, & Jeanneret, 1973, p. 66).

As the final step three groups of either five or six senior college students in personnel were asked to evaluate the 19 benchmark jobs with each group using either the ranking, factor comparison, or point method according to the procedures outlined in a wage salary text (Dunn & Rachel, 1971). The group involved with the factor comparison method was provided with wage and salary survey data on five jobs of their choice from among the benchmark jobs.

Results

The results from application of the various job evaluation methods are reported in Table 1.

When the results from all methods were intercorrelated, coefficients ranged from .82 to .95 with a median coefficient of .89. Thus it was generally found that all methods, including the policy capturing method, yielded similar results.

STUDY 2

Method

In 1973 the job evaluation committee at a major utility in the Intermountain West faced, because of union agreement, the systematic review of 131 classified jobs. Many of the jobs were clerical, craft, or operative and were located in each of the four geographical divisions of the company, although some were found in only one or two divisions. In the judgment of the committee the point method which had been in use yielded generally satisfactory results but required considerable committee time, and in some cases sufficiently objective job information was difficult to obtain. The committee wanted a less costly method which would perpetuate the philosophy of the committee but which would disclose areas of possible inequity.

It was decided to conduct job analyses with the PAQ for the jobs with a minimum of a supervisor and job analyst (an upper-division college student or company personnel officer) independently completing each analysis. Additionally, some analyses were conducted by job incumbents. In cases where the same job was found in more than one division, at least one analysis of a randomly selected position for each job in each division was undertaken. For jobs on which there were substantial numbers of employees, analyses were secured from up to 14 persons.

Once the analyses were completed, the mean values on the PAQ ratings were determined for each job, job dimension (factor) scores were computed, and the mean job dimension scores were compared with the dimension scores from individual analyses of the job following instructions found elsewhere (Mc-

Cormick et al., 1973). This resulted in difference scores (sums of square error) across all dimensions between each individual analysis and the mean analysis. Also, mean and individual scores were plotted as they related to each other. Whenever difference scores exceeded seven, the individual dimension scores known to have had a strong relationship with compensation rates in previous studies (dimensions: B–8, making decisions; D–17, communicating judgments, decisions, and information; and, O–28, having decision making, communication, and social responsibility) were visually inspected. Analyses by incumbents in general proved to be less satisfactory than those done by the personnel department, particularly for blue-collar jobs. Where scores on these dimensions were judged diverse, additional PAQ data were collected to determine whether the positions within the job were actually different or whether differences arose because of analyst error. Additional data were collected on 17 jobs by the personnel department and indicated corrections were made.

Following this reliability and screening process an attempt was made to "capture" the company job evaluation policy. Job dimension scores from the mean analyses for the jobs were used as predictors (in a buildup stepwise regression procedure) of the job evaluation points that had been previously established by the job evaluation committee using a point method. Additionally the predictions were made using the general nonlinear equation (used in Study 1) and were correlated with the company job evaluation points.

Results

Table 2 contains the job dimension scores statistically selected as predictors with their respective weights in the order in which they entered the equation. As indicated in the table, activities involving decision making, communication, and responsibility were positively correlated with compensation rates

TABLE 1

RESULTS FROM FIVE JOB EVALUATION METHODS

Job title	Job evaluation methods				
	Wage and salary survey (median yearly rate for 21 cities) ($)	Policy capturing method (points)[a]	Point method (points)	Factor comparison method (yearly rate)[b]	Ranking method (rank)
Parking ticket clerk	5,296	6,881	193	6,872	5
Meter maid	5,328	7,422	180	5,328[d]	6
Account clerk 1	5,616[c]	6,663	N/A	7,328	7
Library asst., technical processes	5,644	6,334	213	6,477	1
Custodian I	5,828	4,783	120	5,958	2
Park worker	6,094	5,396	100	6,094[d]	3
Caretaker, park dept.	6,418	5,786	113	5,834	4
Animal keeper, head	6,672	10,667	348	11,418	13
Secretary	6,683[c]	5,870	N/A	6,809	8
Laboratory tech./clerk, sewer dept.	8,161	7,147	225	8,241	9
Patrolman	8,210	9,582	330	8,210[d]	10
Park foreman	8,805	9,734	285	10,131	11
Plumbing inspector	9,374	9,713	330	13,078	12
Safety director	9,617	11,584	330	13,203	14
Senior accountant	10,004	11,580	433	14,843	17
City clerk	10,680	10,604	320	10,680[d]	18
Superintendent, sewage treatment plant	10,782	12,822	332	11,958	16
Chief of police	15,580	N/A	403	15,580[d]	15
Director of financial & administrative management	16,224	15,719	493	17,148	19

Note. Data not available is indicated with N/A. All data were collected in 1971.

[a] Converted from monthly to yearly base.

[b] Converted from hourly to yearly base.

[c] From area wage survey (United States Department of Labor, 1970) with an upward cost of living adjustment of 8%.

[d] Supplied from wage and salary survey data.

TABLE 2

JOB DIMENSION SCORES AND WEIGHTS FOUND BY STEPWISE REGRESSION ANALYSIS TO PREDICT COMPANY JOB EVALUATION POINTS

Dimension scores selected as predictors	Regression statistics		
	R at each step of the equation	Simple *r* of predictor with company job evaluation points	B weight[a]
B–8 Making decisions	.82	.82	41.34
F–27 Working on a variable vs. regular schedule	.84	−.26	−7.70
C–14 Using hands and arms to move/position things	.86	−.48	−9.40
O–32 Processing information (an "overall" dimension)	.87	−.26	−17.90
B–9 Processing information (a "division" dimension)	.89	.08	10.04
A–5 Evaluating information from things	.89	.17	6.30
C–11 Using hands and arms to control/modify	.90	.00	−3.29
C–13 Performing activities requiring general body movement	.90	.07	−4.69
O–28 Having decision making/communication/social responsibilities	.90	.75	8.30

Note. The mean of the company job evaluation points for the 131 jobs studied was 227 with a standard deviation of 43 which corresponded to monthly compensation values of $812 and $141, respectively.

[a] Weights given are for equation at nine steps, constant at that step was 227.

while those involving the use of the hands tended to be negatively correlated. The correlation of predictions using the general nonlinear equation with company job evaluation points was .81.

DISCUSSION AND CONCLUSIONS

The results from Studies 1 and 2 strongly support the notion that a policy capturing method (in this case multiple regression applied to PAQ data) can be used successfully to empirically select and weight compensable factors indicative of existing compensation rates. This can be done for jobs found in many companies using various methods of job evaluation or for a single company wishing to perpetuate its own compensation philosophy and policy.

Furthermore, the results indicated that comparable rates could be established with various methods, including the policy capturing method, a finding consistent with the work of others (Chesler, 1948; Davis & Tiffin, 1950; Harris & McCormick, 1973; McCormick, et al., 1972; & L. R. Taylor, personal communication, 1970). Finally, of practical interest was the substantial reduction in cost when the policy-capturing method was used, especially if the PAQ data were also used for establishing selection procedures, classifying jobs, etc., and the relatively objective nature of the data and the evaluation methodology.

REFERENCES

Champagne, J. E., & McCormick, E. J. An investigation of the use of worker-oriented job variables in job evaluation. Unpublished manuscript, Lafayette, Ind.: Purdue University, Occupational Research Center, July 1964. (Prepared for Office of Naval Research under contract Nonr-1100 (19), Report No. 7.)

Chesler, D. J. Reliability and comparability of different job evaluation systems. *Journal of Applied Psychology,* 1948, 32, 465–475.

Davis, M. U., & Tiffin, J. Cross validation of an abbreviated point job evaluation system. *Journal of Applied Psychology,* 1950, 34, 225–228.

Doherty, M. E., & Keeley, S. M. Use of subjective predictors in regression analysis for policy capturing. *Journal of Applied Psychology,* 1972, **56,** 277–278.

Dunn, J. O., & Rachel, F. M. *Wage and salary administration: Total compensation systems.* New York: McGraw-Hill, 1971.

Finn, L. R., & Lee, S. M. Salary equity: Its determination and correlates. *Journal of Applied Psychology,* 1972, **56,** 283–292.

Goldberg, L. R. Diagnosticians versus diagnostic signs: The diagnosis of psychosis from neurosis from the MMPI. *Psychological Monographs,* 1965, 79(9, Whole No. 602).

Goldberg, L. R. Simple models or simple processes? Some research on clinical judgments. *American Psychologist*, 1968, 23, 483–496.

Harris, A. F., & McCormick, E. J. The analysis of rates of naval compensation by the use of a structured job analysis procedure. Unpublished manuscript, Lafayette, Ind.: Purdue University, Occupational Research Center, September 1973. (Prepared for Office of Naval Research under contract N00014-67-A-0266-0016, Report No. 3).

Madden, J. M. *An application to job evaluation of a policy capturing model for analyzing individual and group judgment*. (Tech. Rep. PRL-TDR-63-15, AD-417-273) Lackland Air Force Base, Tex.: Personnel Research Laboratory, Aerospace Medical Division, May 1963.

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. The development and background of the Position Analysis Questionnaire (PAQ). Unpublished manuscript, Lafayette, Ind.: Purdue University, Occupational Research Center, June 1969. (Prepared for Office of Naval Research under contract Nonr-1100 (28), Report No. 5).

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. A study of job characteristics and dimensions as based on the Position Analysis Questionnaire (PAQ). *Journal of Applied Psychology*, 1972, **56**, 347–368.

McCormick, E. J., Mecham, R. C., & Jeanneret, P. R. *User's manual for the Position Analysis Questionnaire: System 1.* West Lafayette, Ind.: University Bookstore, 1973.

Mecham, R. C., & McCormick, E. J. *The use in job evaluation of job elements and job dimensions based on the Position Analysis Questionnaire.* Lafayette, Ind.: Purdue University, Occupational Research Center, June 1969. (NTIS No. AD 691-734) (Prepared for Office of Naval Research under contract Nonr-1100 (28), Report No. 3).

Slovic, P., & Lichtenstein, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, 1971, **6**, 649–744.

United States Department of Labor, Bureau of Employment Security. *Dictionary of Occupational Titles*, Vol. 1, *Definitions of Titles* (3rd ed.) Washington, D.C.: U.S. Government Printing Office, 1965.

United States Department of Labor, Bureau of Employment Security. *Suffix Codes for jobs defined in the Dictionary of Occupational Titles* (3rd ed.) Washington, D.C.: U.S. Government Printing Office, 1967.

United States Department of Labor, Bureau of Labor Statistics. *Area wage survey, the Boise City, Idaho, Metropolitan Area, November 1969.* (Bulletin 1660-34) Washington, D.C.: U.S. Government Printing Office, 1970.

(Received September 17, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 5, 638-642

COMPARISONS OF SELF- AND SUPERIOR RATINGS OF MANAGERIAL PERFORMANCE[1]

HERBERT G. HENEMAN III[2]

University of Wisconsin–Madison

Self- and superior ratings of performance on nine dimensions were obtained from an interorganizational sample of managers by a mail questionnaire. It was found that self-ratings possessed less leniency, restriction of range, and halo error than did superior ratings. The ratings only partially met the requirements of convergent and discriminant validity. Possible explanations for the results and implications for future research on managerial performance were suggested.

Meaningful generalizations about the nature and prediction of managerial performance (Campbell, Dunnette, Lawler, & Weick, 1970; Korman, 1968) are difficult to make. This is due, in part, to the organization and/or criterion-specific nature of most studies.

One research strategy to overcome this deficiency would be to identify performance dimensions common to most managerial jobs, approach managers themselves in different organizations, and obtain self-ratings on the identified performance dimensions. The self-ratings could serve as a substitute for the more difficult to obtain superior ratings. Such procedures would facilitate the design of broad concurrent and predictive validation investigations.

Questions may be raised regarding the potential usefulness of self-ratings, however, especially relative to traditionally obtained superior ratings. For example, a number of studies (Parker, Taylor, Barrett, & Martens, 1959; Prien & Liske, 1962; Thornton, 1968) have found that self-ratings tend to have higher mean values (leniency error) and less variability (restriction or range error) than do superior ratings. Such may not be the case, however, if self-ratings were obtained under conditions in which they could be used only for research purposes (Miner, 1968).

Another important issue pertains to the relative incidence of halo error among self- and superior ratings. While it might be assumed that intercorrelations among performance dimensions would be consistently higher for self-ratings than superior ratings, the evidence suggests the opposite (Lawler, 1967; Parker et al., 1959; Prien & Liske, 1962; Williams & Seiler, 1973).

Finally, the convergent and discriminant validity (Campbell & Fiske, 1959) of self-ratings has been questioned. Regarding convergent validity, Miner (1968) concluded that "while various levels of supervision tend to agree, superior and self ratings rarely do [p. 84]." The evidence (Lawler, 1967; Nealy & Owen, 1970; Parker et al., 1959; Prien & Liske, 1962; Thornton, 1968; Williams & Seiler, 1973), however, is more mixed than this criticism implies.

Little evidence has been obtained on both convergent *and* discriminant validity. Williams and Seiler (1973) found favorable convergent and discriminant validity for self- and superior ratings. Lawler (1967) and Nealy and Owen (1970) found little convergent, and thus discriminant, validity for self- and superior ratings.

In short, the evidence suggests that self-ratings of managerial performance hold promise as a means for expanding the scope of research on managerial performance. The purpose of the present study was to begin work in this direction by (*a*) identifying broad, yet meaningful, dimensions of managerial performance, (*b*) obtaining self- and superior performance ratings for managers employed in different organizations under conditions in which it was apparent that the ratings could be used only for research purposes by the investigator, and (*c*) assessing self- and superior ratings in terms of leniency, restriction of range, halo, and convergent–discriminant validity.

---

[1] Financial support for data collection and analysis was provided by the Wisconsin Alumni Research Foundation through the Graduate School Research Committee, University of Wisconsin–Madison and by the Graduate School of Business, University of Wisconsin–Madison. Support for earlier phases of this research by L. L. Cummings was provided by the Richardson Foundation and the Graduate School of Business, Indiana University.

[2] The author gratefully acknowledges the assistance of Donald P. Schwab and L. L. Cummings.

Requests for reprints should be sent to Herbert G. Heneman III, Graduate School of Business, University of Wisconsin–Madison, 1155 Observatory Drive, Madison, Wisconsin 53706.

METHOD

Performance Measure

A basic prerequisite of managerial performance measurement is the specification of significant dimensions of managerial activity or behavior to be evaluated. In the present study it was necessary to select dimensions which were specific enough to be behaviorally meaningful, yet general enough to be applicable to managerial jobs across organizations. Given this constraint, it was decided that the eight functional dimensions of managerial performance identified by Mahoney, Jerdee, and Carroll (1963, 1964) would be useful (see below).

This decision was based upon several considerations. First, the dimensions theoretically represent activities engaged in, in varying degrees, by managers regardless of specific job or position in the organization. Second, research by Mahoney et al. (1963, 1964) is supportive of the reliability and validity of this taxonomy. For example, a group of industrial supervisors had no difficulty classifying 100 specific performance incidents into the eight dimensions, and there was a median agreement of 70% between their classifications and classifications by the research staff. Mahoney et al. also found high agreement between self-reported estimates of time spent in each activity and time estimates derived from work sampling procedures for two samples of managers. Finally, it was felt that the dimensions, and their descriptions, would possess adequate face validity for the managers involved in the present study.

After specification of the dimensions, a performance evaluation form was developed. An abbreviated behavioral description of each dimension was taken from Mahoney et al. and placed on the form, along with a 7-point rating scale (1 = low performance to 7 = high performance). A ninth dimension and rating scale, pertaining to overall effectiveness, also was included. Performance was thus measured on a total of nine dimensions: (A) Planning, (B) Investigating, (C) Coordinating, (D) Evaluating, (E) Supervising, (F) Staffing, (G) Negotiating, (H) Representing, and (I) Overall Effectiveness.

Data Collection and Sample

The performance evaluation form was incorporated into two questionnaires that were sent to a sample of managers as part of a larger study. In the first questionnaire (self-rating) each manager was asked to rate his own performance on each dimension in his present job. The second questionnaire (superior rating) contained identical performance scales. Each manager was asked to give the questionnaire to his immediate supervisor and request that the supervisor evaluate the manager's performance. Attached to both questionnaires was a letter indicating that responses would be used only for research purposes by the investigator and that individual responses would remain confidential. A self-addressed, stamped envelope was included with each questionnaire so that the manager and the superior could independently return their questionnaires directly to the author.

The two questionnaires were sent to all master of business administration (MBA) graduates of Indiana University during the years 1964–1967 for whom current addresses were known ($n = 218$). A follow-up postcard was sent to those managers who had not returned the questionnaire within two weeks. A telephone call was then made to those managers who had not responded to either the initial or follow-up mail request.

Usable questionnaires were received from 163 MBAs (response rate of 74.7%) and from 121 immediate superiors (response rate of 55.5%). Among these, there were 102 usable MBA-superior pairs of performance ratings. This later group comprised the final sample in the present study. Due to missing data, however, final intercorrelation sample sizes ranged from 45 to 102.

The managers' average salary and length of service in present job were $18,386 and 3.7 years, respectively. The managers had held an average of 1.6 jobs in different organizations since graduation.

RESULTS

To assess relative leniency and restriction of range errors, means and standard deviations of self- and superior ratings were computed and compared (see Table 1). While six mean self-ratings were *lower* than the corresponding superior ratings, suggesting greater leniency error for superior ratings, a sign test[3] suggested this was probably due to chance ($p < .25$). However, three of the self-rating means were significantly lower than the corresponding superior rating means, using a mean differences test for correlated samples (Ferguson, 1959, pp. 138–140).

---

[3] The sign test (Siegel, 1956; pp. 68–75) focuses on the direction of the differences between pairs of observations. Under the null hypothesis, half of the differences are expected to be positive and half to be negative. All subsequent references to the sign test are from the above source.

Six of the nine standard deviations were *larger* for self-ratings than for superior ratings ($p < .25$, sign test). The self-rating variances were significantly greater than the corresponding superior rating variances in three of four instances, however, using a variance differences test for correlated samples (Ferguson, 1959, pp. 142–143).

To assess the relative incidence of halo error, it is necessary to compare intercorrelations among performance dimensions for self- and superior ratings (monomethod-heterotrait triangles in Table 2). There are 36 intercorrelations in each triangle and thus 36 possible comparisons of self-ratings versus superior ratings. The intercorrelation for self-ratings was less than the corresponding intercorrelation for superior ratings in 31 of the 36 comparisons ($p < .001$, sign test) indicating relatively greater halo error for superior ratings. Further evidence of this is provided by comparing the correlations of the eight functional performance dimensions with the overall effectiveness dimension for self- and superior ratings. In all eight instances, the correlation for superior ratings was greater than the corresponding correlation for self-ratings ($p < .008$, sign test). Evidently the superiors were evaluating the managers on a more "global" basis than the managers were evaluating themselves.

Procedures suggested by Campbell and Fiske (1959) were used to assess the convergent and discriminant validity of the ratings. The basic requirement of convergent validity is that correlations between self- and superior ratings on each performance dimension be statistically significant. These correlations are circled in the validity diagonal in Table 2. The validity coefficients were significant for Dimensions A, E, G, H, and I, thus partially fulfilling the convergent validity requirement.

TABLE 1

MEANS AND STANDARD DEVIATIONS OF SELF- AND SUPERIOR RATINGS

Performance dimension	Self-ratings		Superior ratings	
	Mean	*SD*	Mean	*SD*
Planning (A)	5.38	1.10	5.60	.96
Investigating (B)	5.49	1.16	5.95**	.86**
Coordinating (C)	5.56	1.10	5.56	1.12
Evaluating (D)	5.62	.99	5.60	.93
Supervising (E)	5.34	1.19	5.30	1.15
Staffing (F)	4.74	1.26	5.58**	.91*
Negotiating (G)	4.85	1.43	5.49**	1.03*
Representing (H)	5.30	1.17	5.43	1.28
Overall effectiveness (I)	5.53	.79	5.69	.98*

* Difference signcificant at $p < .05$.
** Difference signifiant at $p < .01$.

The first requirement of discriminant validity is that the convergent validity coefficients be larger than their corresponding row and column correlations in the heteromethod-heterotrait triangles. There were 144 such comparisons possible in total, and the convergent validity coefficients were higher in 101 of these comparisons. In addition, a sign test for each performance dimension indicated that the number of times the validity coefficient exceeded the appropriate row–column correlations was statistically significant ($p < .05$) for Dimensions A, B, E, G, H, and I. The first requirement thus was fairly well met in general, and in particular for the dimensions that met the convergent validity requirement.

The second requirement of discriminant validity is that the convergent validity coefficients be larger than their corresponding row and column correlations in the monomethod-heterotrait triangles. In essence, this involves comparisons of convergent validity to halo error. For self-ratings the validity coefficients were larger in 25 of the 72 possible comparisons; a sign test indicated that only for Dimension H was the comparison significant ($p < .05$). For superior ratings the validity coefficients were larger in only 9 of the 72 possible comparisons; a sign test indicated that the comparisons were not significant for any performance dimension. The data thus did not adequately meet the second requirement, especially in the case of superior ratings.

The third requirement of discriminant validity is that the same pattern of intercorrelations should exist in the heterotrait triangles of both the monomethod and heteromethod blocks. This is a very stringent requirement (Lawler, 1967), and it was not met in the present study.

DISCUSSION

Contrary to previous evidence, there was a tendency for self-ratings to be less lenient

TABLE 2

MULTITRAIT-MULTIMETHOD MATRIX FOR MANAGERIAL PERFORMANCE RATINGS

Method		Self-ratings									Superior ratings								
		A	B	C	D	E	F	G	H	I	A	B	C	D	E	F	G	H	I
Self-ratings	A																		
	B	27																	
	C	30	21																
	D	42	06	20															
	E	32	03	18	34														
	F	35	27	42	39	40													
	G	26	20	18	20	17	51												
	H	06	19	19	10	19	10	23											
	I	55	41	39	33	44	36	40	41										
Superior ratings	A	31**	06	12	11	20	17	11	−10	20									
	B	32	18	24	10	14	00	07	−05	29	49								
	C	18	10	17	01	30	18	11	10	21	53	47							
	D	06	06	04	02	08	−04	05	−06	14	48	46	55						
	E	12	02	04	06	39**	−03	07	04	18	53	43	45	44					
	F	34	32	10	10	27	07	18	09	34	38	51	27	25	72				
	G	20	−08	01	17	20	13	32**	12	18	39	25	47	40	53	42			
	H	13	−02	20	18	23	04	06	27*	13	22	26	31	34	16	18	21		
	I	27	07	13	09	38	05	19	08	26**	66	60	67	55	72	67	52	45	

Note. Decimals are omitted; convergent validities are circled; dotted lines represent heterotrait-heteromethod triangles; solid lines represent heterotrait-monomethod triangles. A = Planning; B = Investigating; C = Coordinating; D = Evaluating; E = Supervising; F = Staffing; G = Negotiating; H = Representing; I = Overall Effectiveness.

* $p < .05$.

** $p < .01$.

and more variable than superior ratings in the present study. For three of the performance dimensions (B, F, G), these differences were statistically significant. These results may have been due to the fact that the ratings were obtained under conditions in which the ratings could be used only for research, as opposed to evaluation and decision-making purposes. This explanation should be considered tentative, however, pending research which explicitly investigates the impact of different "use of ratings" conditions on self- and superior ratings. Sharon and Bartlett (1969) found that such conditions had a differential impact on the leniency of subordinate ratings, and it is reasonable to assume that such also would be the case for self- and superior ratings.

Consistent with previous evidence, self-ratings contained less halo error than did superior ratings, which is another factor in favor of self-ratings. The evidence on halo error suggests a possible explanation for Miner's (1968) conclusion that superior-superior ratings have higher convergent validities than do self-superior ratings. If superiors' global assessments of subordinates agree with one another, and these global assessments dominate their evaluations of subordinates on specific performance dimensions, it is likely that the superiors will tend to agree with each other on each performance dimension. On the other hand, since subordinates' self-ratings are more discriminating across dimensions, they are less likely to exhibit high agreement with their superiors' ratings.

The evidence was only partially supportive of the convergent and discriminant validity of the ratings. In part, large halo error in the superior ratings may have inhibited high convergent validity and, in turn, the ability of the superior ratings to meet the second requirement of discriminant validity.

It is also possible that the performance dimensions and descriptions were not very meaningful to the raters. Unfortunately, there is no way of directly assessing this possibility. There are many other possible specifications of broad managerial performance dimensions (see Campbell et al., 1970), however, which may warrant serious consideration in future research.

In summary, the pattern of results in the present study suggests that the possibility of obtaining self-ratings of managerial performance across organizations should be seriously considered in future research on managerial performance. Until additional research is conducted, however, this suggestion may apply only in situations where it is clear that the ratings will be used for research purposes only.

REFERENCES

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, 56, 81–105.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., & Weick, K. E. *Managerial behavior, performance, and effectiveness*. New York: McGraw-Hill, 1970.

Ferguson, G. A. *Statistical analysis in psychology and education*. New York: McGraw-Hill, 1959.

Korman, A. K. The prediction of managerial performance: A review. *Personnel Psychology*, 1968, 21, 295–322.

Lawler, E. E. The multitrait-multirater approach to measuring managerial job performance. *Journal of Applied Psychology*, 1967, 51, 369–381.

Mahoney, T. A., Jerdee, T. H., & Carrol, S. J. *Development of managerial performance: A research approach*. Cincinnati: South-Western, 1963.

Mahoney, T. A., Jerdee, T. H., & Carroll, S. J. The job(s) of management. *Industrial Relations*, 1964, 3, 97–110.

Miner, J. B. Management appraisal: A capsule review and current references. *Business Horizons*, 1968, 11, 83–96.

Nealy, S. M., & Owen, T. W. A multitrait-multimethod analysis of predictors and criteria of nursing performance. *Organizational Behavior and Human Performance*, 1970, 5, 348–365.

Parker, J. W., Taylor, E. K., Barrett, R. S., & Martens, L. Rating scale content: III. Relationships between supervisory- and self-ratings. *Personnel Psychology*, 1959, 12, 49–63.

Prien, E. P., & Liske, R. E. Assessments of higher-level personnel: III. A comparative analysis of supervisor ratings and incumbent self-ratings of job performance. *Personnel Psychology*, 1962, 15, 187–194.

Sharon, A. T., & Bartlett, C. J. Effect of instructional conditions in producing leniency on two types of rating scales. *Personnel Psychology*, 1969, 22, 251–263.

Siegel, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.

Thornton, G. C. The relationship between supervisory- and self-appraisals of executive performance. *Personnel Psychology*, 1968, 21, 441–456.

Williams, W. E., & Sieler, D. A. Relationship between measures of effort and job performance. *Journal of Applied Psychology*, 1973, 57, 49–54.

(Received June 25, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 5, 643–648

EFFECTS OF HEAD RESTRICTION ON DRIVERS' EYE MOVEMENTS AND ERRORS IN SIMULATED DANGEROUS SITUATIONS[1]

JOHN A. ALLEN, JR.,[2] STEPHEN R. SCHROEDER, AND PATRICIA G. BALL

Highway Safety Research Center, University of North Carolina

Two groups of 15 subjects tracked a segment of the Aetna Life Insurance Company training film, *Traffic Strategy*, by manipulating the controls of an Aetna Drivo-trainer station. One group had their heads restricted during the session, the other did not. Restriction of the head resulted in more driving errors and fewer (*a*) long eye movements, (*b*) driving adjustments to critical events to the left of the driver's field of view, and (*c*) fixation errors, especially overshoots. Eye movement frequency was positively correlated with driving adjustments and negatively related to driving errors. It was concluded that (*a*) driving performance is better when the eye–head compensatory system is able to move freely, (*b*) head restriction may bias the distribution of fixation locations, and (*c*) the frequency of shifting fixations may be negatively correlated with driving errors.

In their systems analysis of driver behavior, Smith, Kaplan, and Kao (1970) view the automobile as a rigid, wheeled exoskeleton of the driver. Such a view implies that the systems characteristics of both driver and vehicle be congruent, and, in general, this seems to be true (Schroeder, 1971). Adequate driver performance greatly depends upon good perceptual-motor coordination and synchrony with the processing of information from the driving environment. Synchrony between eye, hand, and foot are critical variables for proper operator performance. Smith and Putz (1970), for example, have shown that coordinated eye and head movements must act in synchrony for accurate tracking and steering behavior. Feedback delays as small as .1 second from steering movements to the oculomotor system were found to sharply decrease subsequent steering adjustments.

Research by Bartz (1966) on humans and more recently Bizzi, Kalil, and Tagliasco (1971) on monkeys has begun to elucidate the complex relationships between eye and head movements during visual scanning. When a stimulus appears in the periphery, the normal sequence of events is a centrating saccade, a following movement of the head, and then a compensatory eye movement which adjusts for traversal of the head so that the stimulus is fixated foveally. Bartz (1966) demonstrated that the initial saccade begins before the head movement and that the compensatory saccades are completed before completion of the head movement. Furthermore, the angle of rotation of the head and that of the compensatory movement always sum to yield the appropriate foveal fixation. This reciprocal relationship holds for a wide range of angular displacements of a stimulus in the visual field.

Bizzi et al. (1971) have replicated Bartz's results by measuring EOGs (electroculograms) from the eye and EMGs (electromyograms) from neck muscles. They observed that during eye–head turning, neural activity reaches the neck muscles before the eye muscles, and that all agonist neck muscles are activated simultaneously regardless of initial head position. Since overt movement of the eyes precedes head movement, the command for the eye–head sequence must be a central neural one, even though it does not specify the serial order.

---

[1] This research was conducted as part of a larger project entitled, Driver License Road Testing (DI-69-001 (002)) contracted with the North Carolina Department of Motor Vehicles and sponsored by the North Carolina Governor's Highway Safety Office.

[2] Requests for reprints should be sent to John A. Allen, Jr., University of North Carolina Highway Safety Research Center, Chapel Hill, North Carolina 27514.

In another experiment (Bizzi et al., 1971), a head restraint with a fast-action brake was used, which allowed the head to be stopped after a saccade was begun but before it was completed. Compensatory eye movements were not necessary in this case and in fact did not occur. It was concluded that compensatory eye movements are not initiated by a central neural command, but are reflex responses resulting from movement of the head, perhaps vestibular in origin.

In driving, a substantial proportion of eye movements are compensatory in nature, yet most eye movement recording systems have restricted the head to some degree (e.g., Mourant & Rockwell, 1970; Whalen, Rockwell, & Mourant, 1968; Zell, 1969). It is possible that in such situations the oculomotor system may display a different search pattern than in those situations where the head moves freely and where the usual eye–head sequence can occur without interference. In situations where only the analysis of the content of fixated material is of interest, head restriction may not be a critical variable. However, head restriction may have important effects when criterion performance involves the relationship of eye movements to vehicular adjustments. The purpose of this study was to examine the effects of head restriction on eye movements and the resulting driving performance.

METHOD

Subjects

Thirty undergraduate volunteers served as subjects as part of their laboratory requirement in introductory psychology at the University of North Carolina. All subjects were licensed drivers with normal uncorrected vision.

Apparatus

The apparatus consisted of a driving simulator situated in a 10 foot × 14 foot room. The simulator was composed of two major components, namely: (*a*) a 30 inch × 40 inch rear projection screen and movie projector and (*b*) an Aetna Drivo-Trainer driving station module (Figure 1). A chin rest was mounted on a cross-bar situated over the subject's lap behind the steering wheel.

Eye movements were recorded by means of a Biometrics, Inc. Model SGH/V-2 Eye Movement Monitor. Eye movement records were displayed and stored on a Technirite T-722 dual-channel chart recorder (Gulton Electronics, Inc.).

A 6-minute, 10-second movie clip entitled *Traffic Strategy* (Aetna Life Insurance Company) was rear projected onto the screen in front of the subject by means of a 16-millimeter Bell and Howell movie projector equipped with an anamorphic lens. Eye-screen viewing distance was 56 inches. Driver control actions (i.e., steering, braking, accelerating, etc.) were recorded on a chart recorder located in an adjacent room. A synchro-tape repeater synchronized events on both chart recorders with the film.

Dangerous Situations

The Aetna film, *Traffic Strategy*, was carefully previewed by judges. With the aid of the training instructions on the sound track of the film it was possible to identify nine dangerous driving situations evenly spaced throughout the film. A dan-

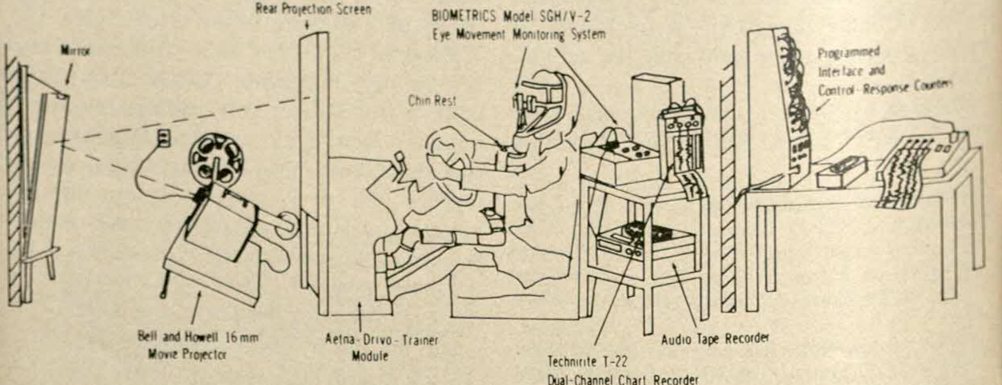


FIGURE 1. Schematic drawing of eye movement laboratory.

gerous event was defined as one in which failure to make a vehicular adjustment would have resulted in a serious accident in the real driving situation.

Judges agreed 100% on the expected vehicular adjustments. The criterion for an acceptable adjustment was simply that it be appropriate to the dangerous situation. For example, if the situation called for the driver to come to a complete stop, decelerating to a low speed without stopping would be considered inappropriate. If the situation called for quickly pulling to the right and the subject responded by lazily turning the car to the right, the response was considered appropriate, although in a real situation it may not have prevented an accident. Errors were defined as failures to make steering, accelerating, decelerating, or braking adjustments within the allotted time. Recording lasted for 20 seconds for each dangerous event since all events lasted 15 seconds or less. Thirty-nine expected vehicular adjustments were distributed among nine dangerous events. Driving errors and total vehicular adjustments per session included those of steering, acceleration, deceleration, and braking movements.

To add realism, traffic sounds from suburban driving were tape-recorded and played through the simulator's speaker. Speaker volume was controlled by the accelerator pedal.

Eye movement response measures included the frequency of long (5-degree visual angle or more) and short (less than 5-degree visual angle) amplitude shifts in fixation and the frequency of fixation errors (both overshoots and undershoots) for each event.

Procedure

When a subject entered the testing room he was seated at the Drivo-Trainer such that his eyes were centered on the screen. A set of tape-recorded instructions followed:

> We are going to show you a short film on this screen. Your job is to track the film as best you can by adjusting the controls of your car. Just drive as you normally would on the highway. First I will calibrate the eye movement recorder, then I will run the film, then recalibrate the recorder.

After adjustment and calibration of the eye movement recorder, the experimenter left the room. The room was darkened during the film. Immediately following the film, recalibration was performed, and the session was ended.

Half of the subjects were run with a head and chin rest, that is, the head restriction group. The other half were run with no head and chin rest, that is, the no head restriction group. The subjects were randomly assigned to one of the two groups.

RESULTS

Driving Adjustments

An analysis of variance was used to compare the total numbers of right and left turns, accelerations, and brake actions of the head restriction and no head restriction groups. There was a significant effect of both head restriction ($F = 11.18$, $df = 1/28$, $p < .005$) and type of vehicular adjustment ($F = 17.64$, $df = 3/84$, $p < .001$). Their interaction was not significant however. As can be seen in Figure 2, the primary significant effects of head restriction were on left turns and accelerations plus decelerations.

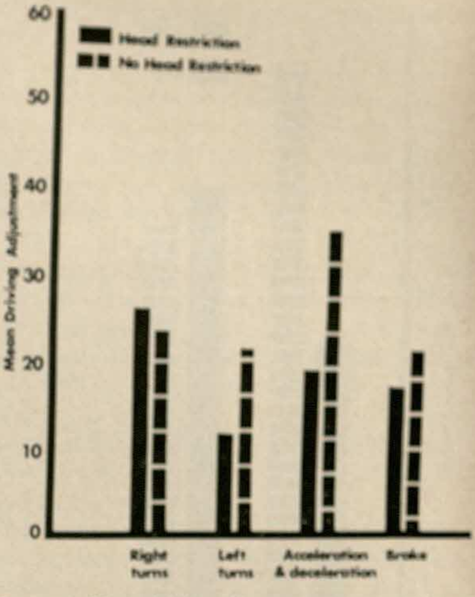


FIGURE 2. Mean driving adjustments as a function of head restriction and type of driving adjustment.

Driving Errors

An analysis of variance of driving errors revealed that errors differed significantly as a function of the type of vehicular adjustment ($F = 40.64$, $df = 3/84$, $p < .001$) but not as a function of head restriction. Once again the interaction between the two variables was not significant.

Eye Movements

An analysis of variance comparing total numbers of eye movements per subject for the head restriction versus no head restriction groups yielded no significant difference. (Mean eye movement frequency per subject was 342 and 353, respectively.)

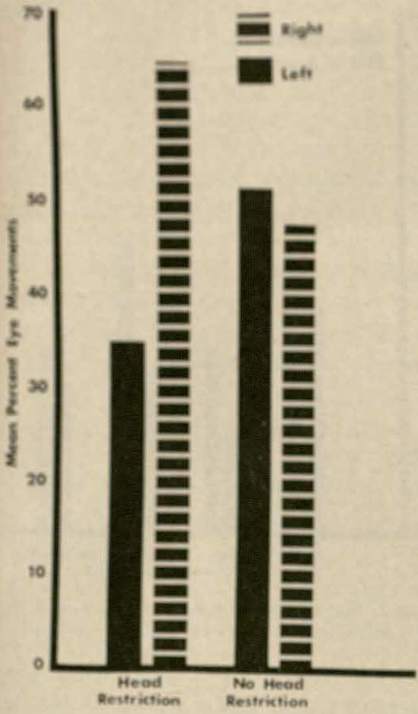


Figure 3. Mean percentage of long excursion eye movements as a function of head restriction.

Since head restriction appeared to have a strong effect on turning left, long left and right excursions of the eye (i.e., 5-degree visual angle or more) were totaled for each group and subjected to an analysis of variance. The main effects of head versus no head restriction and looking right versus looking left were not significant, but their interaction was ($F = 6.5$, $df = 1/28$, $p < .025$). In Figure 3 it can be seen that with head restriction, subjects made proportionally fewer long eye movements to the left than to the right. This difference was highly significant for the head restriction group ($F = 10.30$, $df = 1/28$, $p < .001$) but not for the no head restriction group.

Fixation errors (i.e., undershoots and overshoots) of the two groups were compared by means of an analysis of variance. Although the main effects of head restriction and type of fixation error were not significant, their interaction was ($F = 5.54$, $df = 1/28$, $p < .05$). Head restriction subjects made fewer overshoots ($F = 5.47$, $df = 1/28$, $p < .05$).

Relationship of Eye Movements, Driving Adjustments, and Errors

To examine the relationships between eye movements, driving adjustments, and driving errors, several correlations were computed from the data. For both groups combined, the rank-order correlation between total eye movements and total driving responses for each subject was +.48 ($p < .05$); between total eye movements and driving errors it was −.75 ($p < .001$); and between fixation errors and driving errors it was −.73 ($p < .001$). Since the correlations may have contained artifacts due to combining the two groups, the same correlations were performed on each group separately. Correlations between total eye movements and total driving responses were significant for both groups (head restriction, $\rho = .56$, $p < .05$; no head restriction, $\rho = .47$, $p < .05$). For the head restriction group the correlation between total eye movements and driving errors was −.05, whereas for the no head restriction group it was −.75 ($p < .01$). A similar result was found for driving errors and fixation errors (i.e., head restriction, $\rho = .09$; no head restriction, $\rho = -.71$, $p < .01$).

DISCUSSION

In the present experiment, restriction of the head was associated with an increase in driving errors and a significant decrease in vehicular adjustments (especially accelerations, decelerations, and left turns). This result was not due to a disproportionately small number of critical events occurring to the left of center in the film, since an examination of the film revealed that slightly more dangerous events developed to the left of the driver's line of sight. With respect to driving adjustments, an examination of the records indicated that subjects with no head restraint made many more adjustments than were necessary to successfully negotiate a given event, while subjects with their heads restricted usually made only the minimum necessary, and sometimes less.

Since both groups made nearly the same overall number of eye movements, differences in shifting the eyes from one fixation to another as a function of head restriction cannot be attributed to differences in activation of alertness[1] between the groups. As Figure 3 shows, head restriction was associated with a decrease in long eye movements to the left of the driver's line of sight. This finding complements the fact that subjects with their heads restricted made fewer left turning adjustments than did subjects without restriction. It is doubtful that this result was due to differences in the physical restraint, since head restraint to the left was the same as that to the right.

Although the conditions of the present study were quite different, the results for the head restriction subjects agreed with those obtained in an experiment by Whalen, Rockwell, and Mourant (1968). Using a head-mounted camera in an instrumented car, they found that head restriction subjects tended to concentrate their fixations to the right of the field of view while driving on the open road and in freeway traffic. The data for the no head restriction group did not follow this pattern. It is possible that when the head is restricted subjects tend to concentrate on driving subroutines related to information on the right side of the road (e.g., lane position, car following, road sign information, etc.) at the expense of scanning a broader field of view, as perhaps is more typical of the driver with unrestricted head movements.

The findings with fixation errors support those of Bartz (1966) and Bizzi et al. (1971). Under conditions of no head restriction fixation errors following long excursions were equally distributed between undershoots and overshoots. On the other hand, restriction of the head greatly decreased such movements, especially overshoots.

Since compensatory eye movements were unnecessary in the head restriction group, a question might arise as to why any fixation errors occurred at all. In this regard, Bizzi et al. (1971) had noted that complete elimination of compensatory eye movements occurred only when the location of the stimulus was presented randomly. However, when a target location was not random, such as after practice, some compensatory movements still occurred with head restriction. Given their previous practice, this is likely to be the case with licensed drivers.

The fact that eye movements were found to relate significantly to both driving adjustments and errors emphasizes the close interplay of perceptual-motor systems in driving. The results of the present experiment thus agree with the findings of Smith et al. (1970) and Bizzi et al. (1971), who found that perceptual-motor coordination and the synchrony of eye, hand, and foot are closely related to driving performance.

The results of this study also support those from vigilance studies which focus on the monitoring of static displays and use eye movements as observing responses. Mackworth, Kaplan, and Metlay (1964) and Schroeder and Holland (1968), for example, found that subjects who shifted their fixations more frequently detected many more signals. In the present study, the fact that the close relationship between fixation errors and driving responses with no head restriction did not occur with head restriction suggests that the eye–head compensatory system was disrupted since subjects made fewer overshoots with head restriction than without it.

In summary, restriction of the head was shown to result in a disruption of eye movements and simulator driving performance, especially with regard to accelerator usage and steering movements related to critical events to the left of the driver's field of view. These results have several implications for automobile driving and general highway safety. First, the optimal scanning strategy during driving should be one in which the head remains fixed such that peripheral targets have to be detected by long eye movements alone. Second, eye movements data gathered from recording systems which restrict normal head movements should be interpreted with caution. The present evidence suggests that head restriction may bias the distribution of fixation locations. Finally, rate of shifting fixations may be negatively related to driving errors.

---

[1] Mackworth, Kaplan, and Metlay (1964) have suggested that eye movement rate is an index of alertness.

REFERENCES

Bartz, A. Eye and head movements in peripheral vision: Nature of compensatory eye movements. *Science*, 1966, **152**, 1644–1645.

Bizzi, E., Kalil, R., & Tagliasco, V. Eye-head coordination in monkeys: Evidence for centrally patterned organization. *Science*, 1971, **173**, 452–454.

Mackworth, N., Kaplan, I., & Metlay, G. Eye movement during vigilance. *Perceptual and Motor Skills*, 1964, **18**, 397–402.

Mourant, R. R., & Rockwell, T. H. Mapping eye-movement patterns to the visual scene in driving: An exploratory study. *Journal of Human Factors*, 1970, **12**, 81–87.

Schroeder, S. A. A perspective for research on driver eye movements. In P. Waller (Ed.) *Mushrooming Technology North Carolina Symposium on Highway Safety*, 1971, 4, 87–98.

Schroeder, S., & Holland, J. Operant control of eye movements during human vigilance. *Science*, 1968, **161**, 292–293.

Smith, K., Kaplan, R., & Kao, H. Human factors analysis of driver behavior by experimental systems methods. *Accident Analysis and Prevention*, 1970, **2**, 11–20.

Smith, K., & Putz, V. Feedback factors in steering and tracking behavior. *Journal of Applied Psychology*, 1970, **54**, 183–196.

Whalen, J., Rockwell, I., & Mourant, R. A pilot study of drivers' eye movements. (Tech. Rep. EES 277-1) Columbus, Ohio: Ohio State University, System Research Group, 1968.

Zell, J. K. *Driver eye-movements as a function of driving experience*. (Tech. Rep. IF-16) Columbus, Ohio: Ohio State University, Engineering Experiment Station, 1969.

(Received September 6, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 5, 649-651

SHORT NOTES

INTERACTION OF SUBORDINATE AND LEADER CHARACTERISTICS IN MODERATING THE CONSIDERATION–SATISFACTION RELATIONSHIP[1]

DAVID M. HEROLD[2]

Georgia Institute of Technology

The relationship between leader consideration and worker satisfaction was shown to vary for groups of subordinates reporting (*a*) different levels of need for independence and (*b*) supervisors with different levels of hierarchical influence. In a sample of 91 nonsupervisory industrial workers, relationships between consideration and three facets of satisfaction were significantly positive for those subordinates low in independence who had supervisors low in influence and for subordinates high in independence who had supervisors high in influence. No significant relationships were found for the low-independence–high-influence and high-independence–low-influence groups. No support was found for previous findings that leaders' hierarchical influence alone moderates consideration–satisfaction relationships.

Earlier studies have indicated that the same type of leadership behavior may have different effects on individuals with different personalities (e.g., Sanford, 1950; Vroom, 1960). However, only a limited amount of research has investigated specific personality characteristics of group members, such as degree of dependence–independence, which may interact with specific leadership behaviors. The purpose of the present study was to further examine the possible need to include such subordinate characteristics in leadership research.

Pelz (1952) found that supportive or considerate behavior on the part of influential leaders was related to subordinate satisfaction, while the same behavior exhibited by noninfluential leaders was not. Wager (1965) found that the facilitating effect of influence on the consideration–satisfaction relationship was greatest in groups who, for various reasons, shared an "organizationally marginal" status. Wager (1965) hypothesized that the effect of the marginal status was to make these employees dependent upon others for assistance in attaining their various goals, and therefore, more sensitive to supervisory influence. House, Filley, and Gujarati (1971), studying two research and development departments, found influence to be a moderator of the leader behavior–employee satisfaction relationship in only one of them. The lack of effect in the second group was hypothesized to be a function of Wager's (1965) concept of organizationally marginal status, arguing that people were of high enough status so as not to be dependent on their superiors for need satisfaction.

The hypothesis suggested by some of the contradictory influence findings is that subordinate dependence moderates the effects of specific behaviors of influential leaders on employee satisfaction. In other words, the facilitation of hierarchical influence on the consideration–employee satisfaction relationship (Pelz, 1952) is a function of subordinate independence (House et al., 1971; Wager, 1965). Based on suggestions of past research, we predict a stronger relationship between considerate leader behavior and subordinate satisfaction for leaders who are seen as being influential and subordinates who are dependent, than the relationship between leaders who are seen as being influential and subordinates who are independent. For the low-influence leader no predictions are made other than testing for the effects of different subordinate characteristics.

"Independence," as used in the present study, refers to an individual's reported "need for independence"—an individual characteristic. Neither the Wager (1965) nor the House et al. (1971) study suggested subordinate independence specifically as a personality trait; in fact, they seem to suggest that it is more situation specific than the more pervasive variable investigated in this study. However, some researchers have argued that independence as a personality variable is a relevant subordinate characteristic for purposes of understanding superior-subordinate relationships (e.g., Argyris, 1957; Vroom, 1960). It was, therefore, thought productive to investigate independence in the present context of studying subordinate personality as a moderator of the leader behavior–employee satisfaction relationship.

---

[1] A portion of this article is based on the author's paper presented at the annual meeting of the Eastern Academy of Management, Boston, May 5–6, 1972.

[2] Requests for reprints should be sent to David M. Herold, College of Industrial Management, Georgia Institute of Technology, Atlanta, Georgia 30332.

METHOD

Questionnaire data were collected from 91 nonsupervisory employees of an electronic devices manufacturing company. Questionnaires were administered in the company cafeteria on company time.

Leader consideration scores were obtained using 10 items from Form XII of the Ohio State Leader Behavior Description Questionnaire (Stogdill, 1965). Leader hierarchical influence was measured using a scale of items developed by Comrey, Pfiffner, and High (1954) and Wigdor (1969). The scale measures upward influence of the leader only with respect to matters concerning the personal management of the subordinates and participation in policy decisions. The items used to measure the need for independence (Vroom, 1960) were meant to reflect the individual's predisposition to strive for self-reliance and to do things alone without help. Satisfaction measures were obtained for three facets of satisfaction which were related to areas affected by the hierarchical influence of leaders: (*a*) pay, (*b*) advancement, and (*c*) security. These scales had been previously used by Rizzo, House, and Lirtzman (1970).

TABLE 1

RELATIONSHIP BETWEEN SUPERVISOR CONSIDERATION AND THREE FACETS OF EMPLOYEE SATISFACTION FOR GROUPS WITH HIGH- AND LOW-INFLUENCE LEADERS

Supervisor influence level	Employee satisfaction categories		
	Advancement	Pay	Security
Low hierarchical influence	.25 (39)	.30* (37)	.30* (39)
High hierarchical influence	.31* (45)	.30* (45)	.41** (45)

Note. Numbers in parentheses represent *ns* on which correlation coefficients are based. Respondents for whom data entering the correlations were missing were dropped from the particular analysis accounting for differences between subgroups and total sample size.
$^{*}p < .05$, one-tailed test.
$^{**}p < .01$, one-tailed test.

TABLE 2

RELATIONSHIPS BETWEEN SUPERVISOR CONSIDERATION AND THREE FACETS OF EMPLOYEE SATISFACTION FOR DIFFERENT COMBINATIONS OF LEADER INFLUENCE AND SUBORDINATE INDEPENDENCE

Levels of independence and influence	Employee satisfaction categories		
	Advancement	Pay	Security
Low influence Low independence	.54** (20)	.54** (18)	.46* (20)
Low influence High independence	−.06 (14)	−.09 (14)	.16 (14)
High influence Low independence	−.03 (11)	−.20 (11)	.06 (11)
High influence High independence	.33* (28)	.4[illegible]** (28)	.62** (28)

Note. Numbers in parentheses represent *ns* on which correlation coefficients are based. Respondents for whom data entering the correlations were missing were dropped from the particular analysis accounting for differences between subgroups and total sample size.
$^{*}p < .05$ one-tailed test.
$^{**}p < .01$.

RESULTS

The sample was split at the mean to form high and low supervisor influence groups, and correlations were computed between the supervisor consideration scores and each of the three facets of employee satisfaction.

As can be seen from Table 1, the moderating effect of influence found by past researchers was not supported in the present sample. Relationships between consideration and satisfaction were approximately the same for both the high and low hierarchical influence groups. Most of these relationships were significant.

To test the effects of different subordinate-superior groupings on the aforementioned relationships, the sample was dichotomized on the influence and independence variables simultaneously (the variables were tested and found to be relatively independent of each other so that placement on one variable did not determine an individual's placement on the other). Table 2 reports the correlations between consideration and satisfaction for the resulting groups.

As shown in Table 2, when high and low influence are moderated by subordinates' need for independence, there is a drastic change or qualification of the findings in Table 1. We find strong relationships between consideration and satisfaction for two very different groups (low influence–low independence, high influence–high independence), and no significant relationships between consideration and satisfaction for the other two groups.

DISCUSSION

The prediction made for the high-influence–low-independence group, suggested by previous research, was not confirmed. In fact, the prediction was reversed with the high-influence–high-independence group exhibiting strong positive relationships, while the high-influence–low-independence group exhibited no significant relationships. These results should alert us to the possibility that reported findings of weak or moderate relationships may be masking the existence of two or more population subgroups which exhibit widely different relationships.

The findings of strong consideration–satisfaction relationships for the low-influence–low-independence and the high-influence–high-independence groups seem to suggest a need for further research to better understand the processes which account for these subgroups' congruency. One possible explanation involves a further conceptual development of the independence and influence measures. Perhaps the need for independence also expresses a need to influence one's own life by affecting one's outcomes in a situation. The considerate, upwardly influential, supervisor can act as a facilitator of this self-determination process and thus satisfy his employees. Conversely, the supervisor who lacks upward influence may attempt to exercise greater control or influence downward. If he does this in a considerate manner he may fill a satisfying, paternalistic role for dependent subordinates, while stifling the attempts at self-determination on the part of independent subordinates. While such explanations are obviously post hoc, the findings suggest that such interactions of leader and subordinate characteristics should be more thoroughly investigated and perhaps even replace the unidimensional studies of leader behavior that are then generalized across populations.

Further, contradictory findings in the area of leadership should not necessarily be viewed as the weakening or failure of the construct being studied, but rather, they could be viewed as a failure to sufficiently elaborate the construct or theory to account for different behavior on the part of different population subgroups.

Using the often researched relationship between consideration and satisfaction, we showed this relationship to vary for groups of subordinates having differing levels of need for independence and supervisors described as having different degrees of hierarchical influence. These findings have the potential for resolving the previous contradictory findings pertaining to the leader's influence by demonstrating that the same level of influence on the part of a leader can be accompanied by different relationships between leader behavior and subordinate satisfaction, depending on the characteristics of the subordinates. To the extent that Pelz (1952), Wager (1965), and the House et al. (1971) studies dealt with populations of differing characteristics, they could well have produced seemingly contradictory findings. The elaboration or qualification of some of the present leadership concepts or constructs may be helpful in consolidating this area of abundant but inconsistent findings.

REFERENCES

Argyris, C. *Personality and organization.* New York: Harper, 1957.

Comrey, A. L., Pfiffner, J., & High, W. S. *Factors influencing organizational effectiveness.* Los Angeles: University of Southern California, 1954.

House, R. J., Filley, A. C., & Gujarati, D. M. Leadership style, hierarchical influence, and the satisfaction of subordinate role expectations: A test of Likert's influence proposition. *Journal of Applied Psychology,* 1971, **55**, 422–432.

Pelz, D. C. Influence: A key to effective leadership in the first-line supervisor. *Personnel,* 1952, **29**, 209–217.

Rizzo, J. R., House, R. J., & Lirtzman, S. I. Role conflict and ambiguity in complex organizations. *Administrative Science Quarterly,* 1970, **15**, 150–163.

Sanford, F. H. *Authoritarianism and leadership.* Philadelphia: Institute for Research in Human Relations, 1950.

Stogdill, R. M. *Manual for the Leader Behavior Description Questionnaire, Form XII.* Columbus, Ohio: Ohio State University, Bureau of Business Research, 1965.

Vroom, V. H. *Some personality determinants of the effects of participation.* Englewood Cliffs, N.J.: Prentice-Hall, 1960.

Wager, L. W. Leadership style, hierarchical influence, and supervisory role obligations. *Administrative Science Quarterly*, 1965, **9**, 391–420.

Wigdor, L. Effectiveness of various management and organization characteristics on employee satisfaction and performance as a function of employees need for independence. Unpublished doctoral dissertation, Bernard M. Baruch College, City University of New York, 1969.

(Received September 17, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 5, 652–655

DISPARAGEMENT BY A SUBORDINATE AS AN INFLUENCE ON THE USE OF POWER[1]

EUGENE M. FODOR[2]

Department of Social Sciences, Clarkson College of Technology

The purpose of the present investigation was to determine whether distribution of rewards by student supervisors in an industrial simulation was influenced by a situation in which a member of the work crew repeatedly disparaged their competence. Disparagement did not appear to affect the supervisor's allocation of rewards to an ingratiator, but it did result in a tendency to favor a compliant worker who was not an ingratiator. Under circumstances where he found himself the object of denigration by one of his subordinates, the supervisor gave the compliant worker higher creativity ratings on Tinker Toy models designed during the simulation ($p < .01$) than he did in the control condition of the experiment.

The present investigation was the most recent in a series of experiments by the author dealing with use of power by appointed student supervisors (Fodor, 1973a, 1973b). The general procedure involved an industrial simulation which was suggested in part by studies done by Goodstadt and Kipnis (1970) and Kipnis and Vanderveer (1971).

Under conditions of disparagement by a subordinate, student supervisors were expected in their evaluations to favor more highly both an ingratiating worker and an otherwise compliant worker who was not an ingratiator than they would in the absence of these conditions.

METHOD

Subjects

Experimental participants were men recruited from the general psychology course. All subjects were given course credit for taking part in the experiment, and no one refused to cooperate. Of the subjects, 14% did not believe in the existence of workers in the next room and were consequently disregarded for purposes of data analysis. Sixty subjects remained—30 in the experimental and 30 in the control group.

Procedure

When subjects arrived at the laboratory, they were told that they would act as supervisors in an industrial simulation by directing and attempting to facilitate the efforts of three high school students in the next room whose task it was to create new model designs from Tinker Toys. A different model was to be consructed on each of four, five-minute trials. Student supervisors were informed that the workers had already been thinking about what models they would design for 15 minutes before the beginning of the first trial. To maximize supervisors' sense of personal involvement in the industrial simulation, the experimenter told them that a major purpose of the experiment was to assess their level of managerial ability. It was understood that the workers would each receive a base pay of $1 for their work. A large sign posted in front of the supervisor authorized him to grant each worker a 10¢ increase or decrease in pay at the end of any trial. The three workers were entirely fictitious.

In communicating with workers, the supervisor could say anything he wished; he could give directions, advice on construction, feedback on awards of pay increase or decrease, encouragement, or reprimands.

Supervisors communicated with the workers by means of a one-way communication device comprised of a microphone and also of a switch and light signal for contacting individual workers. Communications allegedly coming from the various workers to the supervisor consisted of preprogrammed, hand-written notes. The experimenter explained that the supervisor could not be allowed to see the workers or to receive voiced communication from them because the sensory impressions of the various workers thereby acquired might bias the supervisor's evaluations.

The model designs attributed to the workers were independently assessed for level of creativity by each of three judges prior to the experiment, and models were distributed across workers in such a way that the mean and exact distribution of creativity scores for each worker, as determined by the judges' ratings, were identical.

---

[1] The research reported herein was supported by Grant MH 21568-01 from the National Institute of Mental Health. Grateful acknowledgement is expressed to Dean Bailey, Thomas Ingram, Francis Sharp, and John Quintana who served as experimenters.

[2] Requests for reprints should be sent to Eugene M. Fodor, Department of Social Sciences, Clarkson College of Technology, Potsdam, New York 13676.

Immediately after each trial, the experimenter (who was described in the initial instructions as the supervisor's assistant) brought to the supervisor the model constructed by each worker on that trial and told him what structures the models were designed to represent. The assistant at that time also gave the supervisor any notes allegedly written by the workers. There were two conditions of the experiment. In the experimental condition, Man A wrote a series of comments which were highly critical of the supervisor. On Trial 1 he said, "These models are tough to put together and you're not very much help." On Trial 2 he declared, "You're a lousy supervisor." On Trial 4, "I hope you're not planning to be a supervisor when you finish college. You're not very good at it."

The comments of Man B were designed to make him an ingratiator in the sense in which that concept has been treated by Jones (1964). On Trial 1, Man B wrote, "I hope you're happy with my work—I'm doing as well as I can." On Trial 3 he commented, "I like the way you supervise—I thought you'd boss me around, but you haven't." On Trial 4, "You've been a good supervisor."

Man C wrote notes which were essentially neutral in character. On Trial 2 he asked, "Can you tell us every so often how much time we have left?" Trial 3 produced the comment, "These tinker toys are always falling on the floor." On Trial 4 he requested permission to smoke.

Comments were so designed that each worker wrote approximately the same total number of words throughout the experiment. Also, the same person always wrote notes for Man A, another for Man B, and another for Man C.

The control condition had Man A, B, and C constructing the same models in the same order as in the experimental condition. Man B made the same ingratiating comments and Man C the same neutral comments. Instead of disparaging the supervisor, however, Man A, like Man C, wrote a series of neutral statements. On Trial 1 he stated, "I've never done anything like this before. It's kind of different." On Trial 2 he commented, "You've got to be pretty careful about these models. They keep falling apart." On the last trial he asked, "This is the last trial, isn't it?" The control condition differed from the experimental only in the absence of any disparagement of the supervisor.

TABLE 1

MEAN PAY INCREASES, PERFORMANCE EVALUATIONS, AND CREATIVITY RATINGS GIVEN COMPLIANT WORKERS

Condition	Evaluation dimension					
	Pay		Performance		Creativity	
	Man B	Man C	Man B	Man C	Man B	Man C
Disparagement						
M	14.0	19.3	24.1[a]	27.7[a]	24.8[b]	28.0[b,c]
SD	9.52	9.80	4.94	3.19	3.72	2.91
Control						
M	15.3	15.0	25.2	25.0	23.1	25.3[c]
SD	10.59	10.09	5.68	5.58	5.08	4.06

[a] These means are significantly different from each other, $p < .05$.

[b] These means are significantly different from each other, $p < .01$.

[c] These means are significantly different from each other, $p < .01$.

Dependent Measures

The supervisor's response to the three workers in the two conditions of the experiment was measured in three different ways. First, he was asked at the end of each trial to rate all models constructed on the degree of creativity he saw reflected in their design on a scale ranging from 1 (low creativity) to 9 (high creativity). Second, he was allowed to give pay increases at the end of each trial. And lastly, at the conclusion of the four work trials the supervisor was led to another room and requested to evaluate the performance of each worker on four 9-point scales: (*a*) worker's ability, (*b*) worker's overall worth to the company, (*c*) willingness to rehire worker for a second experiment, and (*d*) recommendation that the worker be promoted to supervisor in a future experiment. For purposes of analysis, these four scale scores were summated to obtain a single overall score.

Upon completion of the experiment, all subjects were told the true purpose of the experiment, carefully interviewed to determine whether they suspected the nonexistence of workers in the next room, and then pledged to absolute secrecy.

RESULTS

Table 1 shows the mean reward level accorded Man B, the ingratiator, and Man C, the noningratiating compliant worker, in each condition of the experiment. The higher the mean for any of the three dependent measures in the dis-

TABLE 2

MULTIVARIATE AND UNIVARIATE ANALYSES OF VARIANCE

Source	Multivariate *F* value	Univariate analyses			
		df	*F* values		
			Pay	Performance	Creativity
Disparagement vs control condition (*A*)	8.70**	1	.66	.85	8.33**
Difference between compliant workers (B)	14.61**	1	1.83	3.38	14.94**
A × B	2.25	1	2.37	4.15*	.45

$^{*}p < .05.$
$^{**}p < .01.$

paragement as compared with the neutral condition, the greater was the impact of experimental manipulation in influencing supervisory distribution of rewards.

Table 2 presents a multivariate analysis of variance for identifying significant effects of treatment and worker on the three dependent measures taken together. In accordance with the recommendation of Hummel and Sligo (1971), this multivariate analysis is followed by univariate analyses on each separate dependent measure to trace the source of significant multivariate effects. The multivariate analysis discloses significant effects for both treatment ($F = 8.70$, $df = 1/116$, $p < .01$) and worker ($F = 14.61$, $df = 1/116$, $p < .01$). As Table 2 indicates, both main effects can be traced through univariate analyses to the creativity ratings. Table 1 shows that it was Man C, the noningratiating compliant worker, who received a significantly higher creativity rating as a consequence of the disparagement manipulation.

Further inspection of Table 1 reveals a Treatment × Worker interaction on the variable of performance evaluation. The evaluation of Man C in the disparagement condition, as computed by Tukey's HSD (honestly significant difference) test, was not significantly greater than the evaluation he received in the control ($p < .10$), although it was significantly higher than the evaluation given Man B in the disparagement condition ($p < .05$). Table 1 suggests that the latter result arose in part because Man B was evaluated slightly less favorably in the experimental ($\bar{X} = 24.1$) than in the control condition ($\bar{X} = 25.2$).

In no instance did Man B, the ingratiator, receive significantly greater rewards in the disparagement than in the control condition. Indeed, on the variables of pay increase and performance evaluation, he was given slightly less in the disparagement condition. Nor was there any tendency across conditions for the supervisor to bestow greater rewards upon the ingratiator than upon the noningratiating compliant worker.

DISCUSSION AND CONCLUSION

A conclusion seemingly warranted by the present investigation is that when experiencing disparagement by a subordinate, the student supervisor rates as more creative the models designed by Man C, a compliant but noningratiating worker. This finding harmonizes well with the results of the Goodstadt and Kipnis (1970) and the Kipnis and Vanderveer (1971) studies, both of which showed an increase in supervisory dispensation of rewards to compliant workers when a member of the work crew displayed a poor attitude. There was a tendency for Man C also to receive more pay and a higher performance evaluation in the disparagement than in the neutral condition, but these trends did not attain significance. Taken together, these data tentatively suggest that when storm clouds gather about the supervisor it may not be so much the ingratiator who can expect to benefit from the adverse situation but the person who unobtrusively complies with the demands placed upon him.

Kipnis and Vanderveer (1971) found a significant tendency across conditions for student supervisors to favor an ingratiator over otherwise compliant workers who were not ingratiators. A possible explanation for failure of the present study to replicate that finding may have been the nature of the ingratiating comments contained in the notes attributed to Man B. In an endeavor to fully approximate the desired experimental effect they were made ingratiating to a rather blatant degree, substantially more so than the notes used in the Kipnis and Vanderveer experiment. For example, one of the comments in the present study unabashedly stated, "You've been a good supervisor." Ingratiation, if it is highly visible to the target person, may constitute more a source of annoyance and suspicion than of gratification (Jones, 1964).

REFERENCES

Fodor, E. M. Disparagement by a subordinate, ingratiation, and the use of power. *Journal of Psychology*, 1973, **84**, 181–186. (a)

Fodor, E. M. Group stress, ingratiation, and the use of power. *Journal of Social Psychology*, 1973, **91**, 345–346.

Goodstadt, B., & Kipnis, D. Situational influences on the use of power. *Journal of Applied Psychology*, 1970, **54**, 201–207.

Hummel, T. J., & Sligo, J. R. Empirical comparison of univariate and multivariate analysis of variance procedures. *Psychological Bulletin*, 1971, **76**, 49–57.

Jones, E. E. Ingratiation. New York: Appleton-Century-Crofts, 1964.

Kipnis, D., & Vanderveer, R. Ingratiation and the use of power. *Journal of Personality and Social Psychology*, 1971, **17**, 280–286.

(Received October 9, 1973)

Manuscripts Accepted for Publication in the *Journal of Applied Psychology*

Situation-Specific Experience and Nitrogen Narcosis in the Diving Environment. George Moeller (Naval Submarine Medical Research Laboratory, Naval Submarine Medical Center, Groton, Connecticut 06340) and Craig P. Chattin.

Effects of Grouping of Printed Digits on Forced-Paced Manual Entry Performance. E. T. Klemmer (Bell Laboratories, Holmdel, New Jersey 07733) and L. P. Stocker.

Sleep Deprivation Effects on the Vestibular Habituation Process. Patrick J. Dowd (Environmental Physiology Branch, Department of the Air Force, USAF School of Aerospace Medicine (AFSC), Brooks Air Force Base, Texas 78235).

Designing Pre-Employment Training for the Hard to Employ: A Cross-Cultural Psychological Approach. Harry C. Triandis (Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, Illinois 61820), Jack M. Feldman, David E. Weldon, and William M. Harvey.

Perceptual Maps of Supermarket Locations. Richard W. Olshavsky (Graduate School of Business, Indiana University, Bloomington, Indiana 47401), David B. MacKay, and Gerald Sentell.

Effects of Opportunity to Communicate and Visibility of Individual Decisions on Behavior in the Common Interest. Thomas H. Jerdee (Graduate School of Business Administration, University of North Carolina, Chapel Hill, North Carolina 27514) and Benson Rosen.

Utilization of Job Information as a Criterion in Employment Decisions of Interviewers. Yoash Wiener (Department of Psychology, Case Western Reserve University, Cleveland, Ohio 44106) and Mark L. Schneiderman.

Race and Sex as Determinants of Ratings by Potential "Employers" in a Simulated Work Sampling Task. W. Clay Hamner (Graduate School of Business Administration, Michigan State University, East Lansing, Michigan 48824), Jay S. Kim, Lloyd Baird, and William J. Bigoness.

Studies in Interactive Communication: V. Cooperative Problem Solving by Skilled and Unskilled Typists in a Teletypewriter Mode. Gerald D. Weeks, Michael J. Kelly, and Alphonse Chapanis (Department of Psychology, Johns Hopkins University, Baltimore, Maryland 21218).

Developing a Comparative Measure of the Learning Climate in Professional Schools. Donald D. Bowen (Graduate School of Business, University of Pittsburgh, Pittsburgh, Pennsylvania 15260) and Ralph H. Kilmann.

The Effect of the Number of Scale Points on Reliability: A Monte Carlo Approach. Robert W. Lissitz (Department of Psychology, University of Georgia, Athens, Georgia 30602) and Samuel B. Green.

Leadership Orientation as a Function of Interpersonal Need Structure. Charles R. Kuehl (School of Business Administration, University of Missouri, St. Louis, Missouri 63121), Nicholas Di Marco, and Earl W. Wims.

Perceived Leadership Behavior and Employee Confidence in the Leader as Moderated by Job Involvement. Allan P. Jones (Institute of Behavioral Research, Texas Christian University, Fort Worth, Texas 76129), Lawrence R. James, and John R. Bruni.

A Comparative Study of the Work Satisfaction of Females with Full Time Employment and Full Time Housekeeping. Charles N. Weaver (School of Business and Administration, St. Mary's University, San Antonio, Texas 78284) and Sandra L. Holmes.

A Computer Readability Formula Designed for Machine Scoring. Meri Coleman (Department of Psychology, University of California, Davis, California 95616) and T. L. Liau.

The Effects of Supervisors' Sex and Level of Authoritarianism on Evaluation and Reinforcement of Blind and Sighted Workers. Larry D. Baker (School of Business Administration, University of Missouri, St. Louis, Missouri 63121), Nicholas Di Marco, and W. E. Scott, Jr.

Attempting to Operationalize Product Newness. James H. Donnelly, Jr. (College of Business and Economics, University of Kentucky, Lexington, Kentucky 40506) and Michael J. Etzel.

Journal of Applied Psychology
1974, Vol. 59, No. 5, 656–659

PERSONAL CHARACTERISTICS DIFFERENTIATING FEMALE EXECUTIVE FROM FEMALE NONEXECUTIVE PERSONNEL[1]

ROBERT F. MORRISON[2] AND MARIA-LUISE SEBALD[3]

University of Toronto, Canada

A study of personal characteristics differentiating employed executive and nonexecutive women was conducted in a large metropolitan area. Thirty-nine pairs of subjects, matched on age, education, work site, length of employment, etc., were differentiated on level achieved (organization status and policy contribution). The executive group was significantly higher in (*a*) the self-esteem component of need for achievement, (*b*) the need for power, and (*c*) mental ability. Significant differences between the pairs were not found in variables designed to contrast early socialization processes, marital environment, and the need for affiliation.

To provide the opportunity for equal employment to appropriate candidates, it is important for industry to identify and select those women who have the potential for successful executive careers. To aid in implementing this strategy, there is a need for research comparing characteristics of female executives with those of female nonexecutives. The purpose of the present study was to provide such comparisons on characteristics involving early socialization, current environmental factors, and needs.

Hennig (1971) and Epstein (1971), in their discussions of women in business, have emphasized the early socialization process as a factor influencing career aspirations and behavior. On the basis of such earlier studies, it was hypothesized that female executives have significantly different early socialization processes than do female nonexecutives. In particular, it was hypothesized that the executives (*a*) are the only child or eldest in an all girl family; (*b*) perceive themselves more often as the parent's favorite child; (*c*) establish a closer, personal relationship with the father; (*d*) have a father in a professional, business, or executive occupation; (*e*) have parent(s) with a higher educational level; and (*f*) come from first generation, northern European immigrant families.

Bailyn (1964) and Bowman, Worthy, and Greyser (1965) hypothesized that concurrent environmental components such as marital relationships would be important factors. It was hypothesized that the current marital environments of the two groups differ significantly in that more female executives (*a*) are single; and (*b*) when married, receive more support from their husbands in achieving executive status.

A review of pertinent literature dealing with male executives aided in defining the primary factors which differentiated the executive role from that of the nonexecutive (Ghiselli, 1966; Laurent, 1968; McClelland & Winter, 1969). From this review it was hypothesized that female executives have (*a*) a higher need for achievement; (*b*) a higher need for power; (*c*) a lower need for affiliation; and (*d*) a higher level of mental ability, than do the female nonexecutives.

METHOD

Subjects

Information was obtained from 39 pairs of subjects working in a Canadian metropolitan area. Each pair of females consisted of an executive and a nonexecutive matched according to age, education, years of employment, and type of organization. Sampling techniques were adopted to minimize systematic bias so the sample could be considered representative.

Executive subjects had to (*a*) be accountable for a profit center or major functional portion of the organization; (*b*) be responsible for the development and/or approval of organization-wide policy; and (*c*) supervise supervisors. The differences between the executive and nonexecutive groups on the above criteria were significant at or beyond the .01 level using a one-tailed t test.

Measures of Characteristics

The first measure was a three-part biographical questionnaire. Part 1 (11 items) described sample characteristics; Part 2 (55 items) adapted the biographical questionnaires used in the Exxon Corporation study, Early Identification of Management Potential (Laurent, 1968); and Part 3 was developed by the authors to further test the hypotheses (21 items) and to provide unique information (17 items).

---

[1] The authors wish to thank Bell Canada for assisting in the sponsorship of the study.

[2] Requests for reprints should be sent to Robert F. Morrison, Faculty of Management Studies, University of Toronto, Toronto, Ontario, Canada, M5S 1A1.

[3] Now at the Civil Service Commission, Province of Ontario.

The other two instruments were chosen because they had previously discriminated executive from nonexecutive males (Ghiselli, 1966). The Wesman Personnel Classification Test (Form C) was used to measure total, verbal, and numerical mental ability factors, while the Edwards Personal Preference Schedule (EPPS) was used to assess five pertinent individual needs or motives. Pairs of subjects were eliminated from the analysis of the EPPS data when a subject within that pair had a consistency score of eight or less.

To reduce the data and test the judgmental clustering of ordinal and interval items used to assess the hypotheses, a principal-components analysis with an orthogonal rotation to a varimax solution was done. Thirty nondemographic items from the biographical inventory were included in the factor analysis.

RESULTS

Of the 84 variables used to test the hypotheses, 25 yielded significant differences at the alpha level of .05 or better. Only 10 would be expected to occur by chance with probability equal to .01 (Sakoda, Cohen, & Beall, 1954). The factor analysis of biographical items yielded six factors (see Table 1), accounting for 42.5% of the variance, of which five were consistent with rationally developed clusters previously designed to test the hypotheses. These six factors were titled

TABLE 1

BIOGRAPHICAL QUESTIONNAIRE (BQ) FACTORS DESCRIBING EXECUTIVE AND NONEXECUTIVE WOMEN

Factor	Representative BQ Items	t^a
Leadership (15.7% of variance)	Undergraduate organizations	−1.41
	More undergraduate leadership jobs	−2.16*
	More high school leadership jobs	−2.11*
	More high school organizations	−2.04*
Self-esteem (7.2% of variance)	Rates self high as a supervisor	3.82***
	Has exploited opportunities	2.18*
	Most effective in administrative job	2.30**
	Sways group of associates	2.85**
Supervisory Role Perceptions (5.5.% of variance)	Desires to work at top levels	5.71***
	Gives subordinates more chances	−4.00***
	Supervises more employees	−3.87**
	Fewer lunches with subordinates	−3.16*
Early Career Socialization Process (5% of variance)	Many teen physical activities	−0.10
	Exclusive early home	1.58
	Parents encouraged college	−0.23
	Discussed career with parents	−1.75
Social Skills (4.9% of variance)	Others contact her with problems	−1.42
	Expresses views in groups	2.85**
	Gets along well with others	1.32
	Turns down subordinates lunches	−3.16*
Independence (4.2% of variance)	Higher level of education	−0.78
	Paid more college costs	−0.09
	Too much contact with boss	−1.57
	Less sociable	1.32

[a] $df = 38$; one-tailed for hypothesized variables.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

Early Career Socialization Process, Self-esteem, Independence, Leadership, Supervisory Role Perceptions, and Social Skills. Individual items from within a factor were analyzed for significance between the two groups without the development of factor scores.

The Early Career Socialization Process factor was composed of 10 items primarily assessing early family life and socioeconomic status. Representative items are shown in Table 1. No items on this factor differentiated between the two groups. Other related nominal items which could not be included in the factor analysis also did not discriminate between the two groups. Additional items evaluating early family life included information about the number and ordinal position of siblings, relationships with parents, and intra- and extra-family relationships. Socioeconomic status items included contrasts such as immigrant status, ethnic background, parental education and occupation, and family income.

Atkinson (1958) has stated that high self-esteem and independence are characteristics of those individuals with need for achievement. A second factor, Self-esteem, discriminated between the two groups. In addition to the items in Table 1, executives felt they had achieved more than their educational peers, had more good breaks, and desired administrative work.

Independence was another characteristic which was present in the factor analysis. None of 10 items similar to those in Table 1 discriminated between the groups.

Of the additional nine biographical questionnaire items used to measure need for achievement (need to complete work, take median level risks, plan for the future, etc.), only "the desire to assume responsibility" demonstrated a significant difference between the two groups with executives requesting more. While the Achievement scale on the EPPS supported ($t = 1.78$, $p < .05$) the hypothesis that female executives have a higher need for achievement, it appears that only the characteristic of self-esteem within the overall need for achievement measure discriminated between the two groups. No real support was found for the hypothesis about need for achievement.

The need for power (Atkinson, 1958) was measured through the EPPS Dominance and Deference scores as well as 19 biographical data items. As hypothesized, the Dominance score on the EPPS was significantly ($t = -3.14$, $p < .001$) higher for executives, but the Deference score was not higher for the nonexecutives.

Two of the factors, Leadership and Supervisory Role Perceptions, may be considered as representing a need for power. The six items with major loadings on the factor titled Leadership were leadership position and group membership items (Table 1). Five items loaded onto Supervisory Role Perception and discriminated significantly between the two groups (Table 1).

Thus, it appeared that there was indeed a difference between female executives and nonexecutives regarding the need for power, with executives having a higher need than did nonexecutives.

Two scales from the EPPS, Affiliation and Nurturance, appeared to measure the characteristics of need for affiliation as defined by Atkinson (1958). Using a one-tailed t test, both showed significant differences ($p < .05$) in the hypothesized direction with executives lower on each. Of the 19 biographical data items hypothesized to discriminate between the two groups, only four did so. Though all items were in the hypothesized direction, such results are not meaningful (Sakoda et al., 1954), and the hypothesis is not accepted.

A clear cluster representing the need for affiliation did not come out of the factor analysis which had included six rationally developed measures of need for affiliation. The Social Skills factor contained elements of need for affiliation combined with social influence items from the need for power (Table 1). This indicates that social skills reflect the ability to relate to people as required to fulfill the need for power rather than the need for affiliation.

Using a one-tailed t test, all parts of the Wesman test, verbal, quantitative, and total scores, substantiated ($p < .01$) the hypothesis indicating that the female executives illustrated more intellectual capabilities.

The husband's role behavior (5 items) in supporting his wife's career, career aspirations, business problems, and potential geographic relocation did not differentiate significantly between the two groups.

DISCUSSION

Since the hypotheses involving the need for power, mental ability, and a major element (self-esteem) within the need for achievement were supported in the research and had been developed from research on male executives, it may be possible that the female executive is similar in motivation and ability to the male executive. Cross-validation and further research should be conducted to adequately assess such a conclusion.

Rather than assume that the socialization process is not an important factor in female executive success, future research might consider that early socialization may have a major impact on occupational choice between a career in an organization or in a home. Later experiences within the organization (Berlew & Hall, 1966) may affect the individual's choice about, as well as progress toward, an executive position.

REFERENCES

Atkinson, J. W. (Ed.) *Motives in fantasy, action, and society*. Princeton, N.J.: Van Nostrand, 1958.

Bailyn, L. Notes on the role of choice in the psychology of professional women. *Daedalus*, 1964, **93** (Spring), 700–710.

Berlew, D. D., & Hall, D. T. The socialization of managers: Effects of expectations on performance. *Administrative Science Quarterly*, 1966, **11**, 207–223.

Bowman, G. W., Worthy, N. B., & Greyser, S. A. Are women executives people? *Harvard Business Review*, 1965, **43**(4), 14–28, 164–178.

Epstein, C. F. *Women's place, options and limits in professional careers*. Berkeley and Los Angeles, Calif.: University of California Press, 1971.

Ghiselli, E. E. *The validity of occupational aptitude tests*. New York: Wiley, 1966.

Hennig, M. What happens on the way up? *M.B.A.*, 1971, **5**(6), 8–10.

Laurent, H. Research on the identification of management potential. In J. A. Myers, Jr. (Ed.), *Predicting managerial success*. Ann Arbor, Mich.: Foundation for Research on Human Behavior, 1968.

McClelland, D. C., & Winter, D. G. *Motivating economic achievement*. London: Collier-Macmillan, 1969.

Sakoda, J. M., Cohen, B. H., & Beall, G. Test of significance for a series of statistical tests. *Psychological Bulletin*, 1954, **51**, 172–175.

(Received August 30, 1973)

ANNOUNCEMENT

As of June 1, 1974 the editorial office of the *Journal of Applied Psychology* moved to a new address. Please send all manuscripts and correspondence regarding manuscripts to Dr. Edwin A. Fleishman, Editor, American Instiutes for Research, 3301 New Mexico Avenue, N.W., Washington, D.C. 20016.

Journal of Applied Psychology
1974, Vol. 59, No. 5, 660–662

UNIVERSITY CLIMATE PERCEPTIONS BY BLACK AND WHITE STUDENTS[1]

C. MICHAEL PFEIFER, JR.[2] AND BENJAMIN SCHNEIDER

University of Maryland

Differences in black student and white student perceptions of university climate were examined at an integrated university. Factor analyses of a 115-item university climate questionnaire were performed in racial subgroups. Five factors were extracted for whites and six for blacks. Two racism factors appeared in the black sample, while only one racism factor appeared in the white sample. There were consistent and significant differences on the factor scales between the two races, with blacks perceiving the university climate more negatively.

In recent years much effort has been expended in the measurement of university climate (Feldman, 1971). Previous studies attempting to describe college environments have been oriented to interinstitution comparisons (Astin & Holland, 1961; Stern, 1970). Such comparisons are meaningful in evaluating overall institutional effects on college students as a group. However, it seems equally important to understand the perceptions of university climate in different subgroups within the university.

The little research that has been performed comparing climate-type variables for blacks and whites suggests that people in different groups at the same university may function under different conditions and perceive the climate differently. Hedegard and Brown (1969) found differences in perceived academic competition, friendliness of personnel, participation in activities, and satisfaction with college. Willie and Levy (1972) noted a separatism between blacks and whites on integrated campuses that could lead to different university-related experiences. The purpose of the present research was to investigate intrainstitution differences in perceived university climate for black undergraduate and white undergraduate students.

METHOD

Forty-three black students and 73 white students at the University of Maryland were asked to write essays describing the university. Four different sets of instructions were used to introduce the essay-writing task. Half of the essay sessions were proctored by a white experimenter and half by a black experimenter. These essays and follow-up interviews with 20 black students and 17 white students were used to generate item content for the university climate questionnaire.

From an initial pool of 294 items describing university life, 115 were chosen for the final form of the questionnaire. A total of 730 white students and 138 black students were asked to indicate the extent of their agreement with each of the 115 items on a 5-point scale.[3] Subjects were recruited from an introductory psychology course, higher level psychology courses, Afro-American Studies Program courses, and an Intensive Educational Development program.

The items of the questionnaire were subjected to a factor analysis in separate racial samples. All 138 black respondents were used in the black factor sample. In the white factor sample 550 white respondents were used leaving 180 respondents for a white holdout sample. A principal-factor solution was used (BMDX72) with maximum row correlations as communality estimates. An orthogonal rotation was performed using the varimax method (Kaiser, 1956).

Five factors were rotated in the white sample, and six factors were rotated in the black sample. An eigenvalue cutoff of 2.0 was adopted when deciding the number of factors to be rotated, although fewer factors were rotated in the black sample because of the presence of weak factors which accounted for very little variance. An item was retained for scoring if its loading was .35 or greater in absolute value on a given factor and no higher than .25 in absolute value on any other factor. However, items which loaded .50 or greater in absolute value were retained regardless of their loadings on other factors.[4]

---

[1] The authors wish to thank Duncan L. Dieterly, C. J. Bartlett and Dorothy Evans for their comments and suggestions. Computer time was provided by the Computer Science Center, University of Maryland.

[2] Reprint requests should be sent to C. Michael Pfeifer, Jr., who is now associated with Westinghouse Behavioral Safety Center, Box 948, American City Building, Columbia, Maryland 21044.

[3] Copies of the climate questionnaire are available from the first author.

Factor scores were computed by summing responses to items loading on each factor and dividing by the number of items responded to.

RESULTS

Five factors which emerged in the white sample, along with item examples, were as follows (factor loadings are in parentheses):

Impersonal Academic Atmosphere—a view of the academic process as inadequate, for example, (*a*) "Courses at the University are designed more to flunk people out than to teach them" (.58); and (*b*) "Faculty members at Maryland expect too much from their students" (.54).

Administrative Neglect—a lack of concern for students on the part of the administration, for example, (*a*) "The administration has been responsive to students' needs" (−.60); and (*b*) "The University's administration is behind the times" (.59).

Social Interaction—an atmosphere conducive to meeting and talking with other students, for example, (*a*) "The University of Maryland teaches you to get along with other people" (.55); and (*b*) "A person has the opportunity to meet others with his own ideals at the University" (.52).

Racism—dealing directly with racial matters, for example, (*a*) "Maryland's new admissions policy was devised to keep blacks out" (.59); and (*b*) "Athletic ability is the only thing blacks are praised for at the University" (.52).

Racial Separatism—a lack of interaction between races at the group level, for example, (*a*) "Whites at the University feel uneasy when blacks get together in groups" (.57); and (*b*) "Black students at the University do not trust white students" (.49).

Six factors identified in the black sample were as follows:

Institutional Racism—concerning matters such as admissions standards and the motives of the administration in admitting blacks, for example, (*a*) "Maryland's new admissions policy was devised to keep blacks out" (.68); and (*b*) "The only reason the University admits blacks is to get Federal money" (.56).

Nonacademic Atmosphere—concerning the learning process, for example, (*a*) "The faculty at the University create an atmosphere for learning" (−.54); and (*b*) "There is too much red tape at the University" (.53).

Social Isolation—indicating difficulty in meeting and talking with people, for example, (*a*) "It is difficult to make new friends at the University" (.54); and (*b*) "Maryland's campus is so formal that you feel you can't relax" (.46).

Personal Racism—dealing with interactions between black and white students at the individual or group level, for example, (*a*) "Whites at the University feel superior to blacks" (.56); and (*b*) "Blacks are unwanted in fraternities or sororities at Maryland" (.54).

Nonclassroom Related Activities—concerning such varied things as sports, remedial programs, geographic location, extracurricular activities, for example, (*a*) "Remedial programs on campus help a lot of students" (.46); and (*b*) "Sports are good for the University climate" (.45).

Attempts at Communication—concerning communication between students generally, between black and white students specifically, and between students and the administration, for example, (*a*) "Students at the University create an easy-going atmosphere for each other" (.41); and (*b*) "Whites at Maryland force themselves to be nice to blacks even if they are prejudiced" (.41).

Table 1 presents means, standard deviations, and internal consistency reliability estimates for each factor. Black students perceived the University more negatively than did white students ($p < .01$) on all but one of the climate scales, Attempt at Communication.

DISCUSSION

One of the main differences in climate factors between the black sample and white sample was the emergence of two racism factors in the black sample. (The one racism factor which appeared in the white sample contained items from both of the factors of the black sample.) An important implication of these results is that black students perceive differences between what happens to blacks on a person-to-person level (Personal Racism) and what the institution does to blacks as a class of people (Institutional Racism). Gurin, Gurin, Lao, and Beattie (1969) found a similar distinction among blacks between

---

[4] Tabulations of the items on each rotated factor and their factor loadings are available upon request. Tables of intercorrelations of factors are also available.

TABLE 1

MEANS, STANDARD DEVIATIONS, AND SIGNIFICANCE OF MEAN DIFFERENCES

Variables	Sample	*M*	*SD*	*N*	Internal Consistency Reliability (KR-20)
White sample factors					
Impersonal Academic Atmosphere	White	3.21	.55	180	.71
	Black	3.76	.49	138	.67
Administrative Neglect	White	3.71	.62	180	.72
	Black	3.96	.58	138	.60
Social Interaction	White	3.58	.59	180	.77
	Black	3.20	.54	138	.68
Racism	White	2.26	.61	180	.73
	Black	3.92	.62	138	.71
Racial Separatism	White	3.23	.65	180	.61
	Black	3.55	.50	138	.25
Black sample factors					
Institutional Racism	White	2.67	.52	180	.68
	Black	4.06	.57	138	.77
Nonacademic Atmosphere	White	3.33	.56	180	.51
	Black	3.53	.61	138	.59
Social Isolation	White	3.07	.56	180	.70
	Black	3.28	.57	138	.72
Personal Racism	White	3.28	.46	180	.49
	Black	3.82	.57	138	.71
Nonclassroom	White	3.71	.53	180	.41
	Black	3.55	.55	138	.54
Attempt at Communication	White	3.12	.51	180	.09
	Black	3.12	.69	138	.47

Note. For each pair differences between the means of whites and blacks are significant ($p < .01$), except for Attempt at Communication.

personal items and ideological items when factor-analyzing items measuring Rotter's concept of internal–external control (Rotter, 1966).

Another difference which emerged between the two groups involved the perception of the social climate. Although a factor dealing with social climate emerged in both groups, there were differences in the items loading on the factor. Items loading on the Social Interaction factor of the white sample were, for the most part, positively worded, while items on the Social Isolation scale of the black sample were negatively worded.

In addition to differences in dimensionality or factor structure of university climate for blacks and whites, a comparison of the rank order of means indicated that blacks and whites perceive the behavioral conditions these factors represent as existing in different degrees. The rank-order correlation of the means for blacks and whites was −.85. This difference in rank order might be a function of the salience of the dimensions for the two groups.

Future research might find it useful to examine the ways in which different subgroups of people use the cues available in the environment to structure their perceptions of the world around them. The present data suggest that when the set of cues may be of particular salience to a group they will make finer discriminations among them (e.g., racism cues for blacks) than when the cues are from a set of events not so meaningful or important (e.g., racism cues for whites).

To summarize, dimensions of university climate appear to differ for black students and white students. In addition, blacks consistently perceived the university in a more negative way. Since some research suggests a link between climate and academic performance, future research investigating university climate may help to insure the equality of the total educational experience, including academic performance, for different races. By identifying conditions and situations within the university which are of particular salience for subgroups of students, changes in university policies and procedures can be effected.

REFERENCES

Astin, A. W., & Holland, J. L. The environmental assessment technique: A way to measure college environments. *Journal of Educational Psychology*, 1961, **52**, 308–316.

Feldman, K. A. Measuring college environments: Some uses of path-analysis. *American Educational Research Journal*, 1971, **8**, 51–70.

Gurin, P., Gurin, G., Lao, R., & Beattie, M. Internal–external control in the motivational dynamics of Negro youth. *Journal of Social Issues*, 1969, **50**, 135–142.

Hedegard, J. M., & Brown, D. R. Encounters of freshmen with a public multiversity. *Journal of Social Issues*, 1969, **25**, 131–144.

Kaiser, H. F. The varimax method of factor analysis. Unpublished doctoral dissertation, University of California, 1956.

Rotter, J. B. Generalized expectancy for internal vs. external control reinforcement. *Psychological Monographs*, 1966, **80**(1, Whole No. 609).

Stern, G. G. *People in context*. New York: Wiley, 1970.

Willie, C. V., & Levy, J. D. On white campuses, black students retreat into separatism. *Psychology Today*, March 1972.

(Received September 6, 1973)

JOURNAL OF APPLIED PSYCHOLOGY

Copyright © 1974 by the American Psychological Association, Inc.

December 1974 Vol. 59, No. 6

Diary No. 874
31/3/75

Notice to APA Authors 663

ARTICLES

Studies in Interactive Communication: V. Cooperative Problem Solving by Skilled and Unskilled Typists in a Teletypewriter Mode
Gerald D. Weeks, Michael J. Kelly, and Alphonse Chapanis 665

Effects of Grouping of Printed Digits on Forced-Paced Manual Entry Performance
E. T. Klemmer and L. P. Stocker 675

Effects of Readability of Context Passages on Subsequent Inspection Rates
Ernst Z. Rothkopf and Richard P. Coatney 679

Formats for Presenting Procedural Instructions *Arthur S. Blaiwes* 683

Designing Preemployment Training for the Hard to Employ: A Cross-Cultural Psychological Approach *Harry C. Triandis, Jack M. Feldman, David E. Weldon, and William M. Harvey* 687

Behavior Modification and Absenteeism: Intervention in One Industrial Setting
Ed Pedalino and Victor U. Gamboa 694

Use of Job Information as a Criterion in Employment Decisions of Interviewers
Yoash Wiener and Mark L. Schneiderman 699

Race and Sex as Determinants of Ratings by Potential Employers in a Simulated Work-Sampling Task *W. Clay Hamner, Jay S. Kim, Lloyd Baird, and William J. Bigoness* 705

Effects of Opportunity to Communicate and Visibility of Individual Decisions on Behavior in the Common Interest *Thomas H. Jerdee and Benson Rosen* 712

Effects of the Shorter Workweek on Selected Satisfaction and Performance Measures
John M. Ivancevich 717

Self-Esteem, Performance, and Satisfaction: Some Tests of a Theory
Jeffrey H. Greenhaus and Irwin J. Badin 722

Note on the Utility of Certainty Weights in Expectancy Theory *Jack M. Feldman* 727

An Analysis of Situational Moderating Effects on the Relationships Between Least Preferred Co-Worker and Other Psychological Measures
Marshall Sashkin, F. Carter Taylor, and Rama C. Tripathi 731

Some Results with the German Form of the Position Analysis Questionnaire (PAQ)
E. Frieling, W. Kannheiser, and R. Lindberg 741

Sleep Deprivation Effects on the Vestibular Habituation Process *Patrick J. Dowd* 748

Integration of Attributes in Bus Transportation: Two Modeling Approaches
Kent L. Norman and Jordan J. Louviere 753

Relationship Between Consumers' Category Width and Trial of New Products: A Reappraisal *Lyman E. Ostlund and Brynulf Tellefsen* 759

(*Continued on next page*)

(Continued from previous page)

Attempting to Operationalize Product Newness: A Reply to Ostlund and Tellefsen
James H. Donnelly, Jr., and Michael J. Etzel 761

SHORT NOTES

Using Subordinate Ratings to Elicit Behavioral Changes in Supervisors
W. Harvey Hegarty 764

Cross-Sample Stability of the Supervisory Behavior Description Questionnaire
Andrew D. Szilagyi and Henry P. Sims, Jr. 767

Relation Between Foremen's Leadership Attitudes and the Skill Level of Their Work Groups
Chris Ching-Yang Hsu and Richard R. Newton 771

Mechano-Content Analysis as an Aid to Predicting Effects of Advertising Microcommunications
Donald W. Hendon 773

Effects of Alternative Positioning of Open-ended Questions in Multiple-Choice Questionnaires
W. Russell Johnson, Nicholas A. Sieveking, and Earl S. Clanton III 776

LIST OF MANUSCRIPTS ACCEPTED 763

SPECIAL ANNOUNCEMENT 766

INSTRUCTIONS TO AUTHORS

Articles must possess: (*a*) conciseness; (*b*) unambiguous and simple vocabulary; (*c*) conformity to accepted technical style in tables, terminology, and references; and (*d*) conclusions that are clearly related to the evidence presented. Authors must progress from a statement of problem or purpose, through analysis of evidence, to conclusions and implications.

Criteria for evaluation include: (*a*) significance of contribution, (*b*) technical adequacy, (*c*) appropriateness for the *Journal of Applied Psychology,* and (*d*) clarity of presentation.

See the *Publication Manual of the American Psychological Association* (2nd ed.), $3, for the format of manuscripts. Articles not prepared in this manner cannot be reviewed. Note sections on typing (*all copy must be double-spaced*), tables, figures, and references. Do not abbreviate, for example, journal titles in reference lists, the word *subject,* names of groups, etc. Each manuscript copy must include an abstract of 100–120 words typed on a separate sheet of paper.

Optional blind review must be specifically requested when a manuscript is first submitted. The author's name and affiliation must appear only on a separate title page, included with each copy of the manuscript. Footnotes containing the identity of the author or his affiliation must be on separate pages.

Check recent issues of the *Journal of Applied Psychology* for approximate length of regular articles. (One printed page equals three double-spaced manuscript pages.) A few longer articles of special significance are printed as monographs. Short Notes feature brief reports on studies such as those involving some methodological contribution or important replication.

The original drawing of a figure or an 8 × 10 inch glossy print must be submitted with the original copy of the manuscript. Duplicate copies may be photographic or pencil drawn. Figures *must* be hand-lettered.

Supplementary materials formerly deposited with the National Auxiliary Publications Service need no longer be submitted with the manuscript. Retain supporting and raw data for at least 5 years after publication of the article.

Notice to APA Authors

The American Psychological Association announces publication of the second edition of the *Publication Manual* in August 1974. This new edition, which supersedes the 1967 Manual, will be adopted by the 14 APA journals in 1975.

The new Manual is more comprehensive than the previous edition. It updates APA publication policies and procedures and incorporates changes in editorial practice since 1967. For instance, APA now sends many authors their copy-edited manuscripts for review before they are set into type, and some editors now use blind review procedures. The second edition also includes up-to-date statements of the coverage of each APA journal including the *Journal of Experimental Psychology* which will be published in four separate sections in 1975.

The new *Publication Manual* initiates several changes in APA style. These changes are announced in the August 1974 *American Psychologist* and will be introduced in the APA journals in January 1975. During the period of transition to the new style, authors should note that (*a*) all manuscripts published in 1974 will be copy-edited according to the 1967 Manual, (*b*) manuscripts accepted in 1974 and published in 1975 will be copy-edited to conform to the new Manual. Starting in 1975, accepted manuscripts that depart significantly from the Manual will be returned to authors for correction.

Authors will be encouraged by the changes in the second edition. The new APA style simplifies reference forms; eliminates unnecessary underlines, brackets, and other devices; supports appropriate use of "I" and "we"; and generally clarifies typing requirements. Material is arranged for maximum convenience to authors and typists, and all sections are cross-referenced and indexed.

The new *Publication Manual* is available after August 1 for $3. Send orders to APA Publication Sales, 1200 Seventeenth Street, N.W., Washington, D. C. 20036. Orders of $15 or less must include payment unless they are submitted on institutional purchase order forms.

Journal of Applied Psychology
1974, Vol. 59, No. 6, 665–674

STUDIES IN INTERACTIVE COMMUNICATION:
V. COOPERATIVE PROBLEM SOLVING BY SKILLED AND UNSKILLED TYPISTS IN A TELETYPEWRITER MODE[1]

GERALD D. WEEKS, MICHAEL J. KELLY, AND ALPHONSE CHAPANIS[2]

Johns Hopkins University

Two-person teams of skilled and of unskilled typists ($N = 16$) cooperated to solve credible, real-world problems by communicating through interconnected typewriters or, as a control, by conversing face-to-face. Performance was assessed by (a) the time taken to solve a problem, (b) measures of behavioral activity, and (c) several measures of verbal output. All three criteria showed large differences between face-to-face conversation and communication via typewriter. There were, however, no differences between the two groups of typists on any criteria even when they communicated by typewriter, in part because one-third or less of the total time was spent typing, and in part because subjects composed their own messages. These findings have implications for the use of keyboard terminals in telecommunications and interactive computer systems.

The typewriter keyboard is a common intermediary in communication between people and virtually the only widely used medium of interaction between men and computers. According to Michaels (1971), there are approximately 45 million typewriters in the United States. One reason the typewriter is so popular is that it can be used by people with a wide range of skill levels. A secretary with training and experience may reach a typing rate of 80, 100, or even more words per minute. On the other hand, a person with no training can sit at a keyboard and do a respectable job of pecking out a personal letter or even an article such as this.

Interactive keyboard terminals in time-shared computer systems are already common in a wide variety of business, industrial, and educational settings. The continued proliferation of these systems and their possible future evolution into computer utilities (see, for example, Irwin, 1966) are certain to result in an increase in the diversity of both their applications and the kinds of people who use them.

The high costs associated with the use of computer terminals and the value of a user's own time have prompted wide-scale efforts to increase the efficiency of man–computer interactions. Since users of both typewriters and terminals come to their tasks with a wide range of typing skills, one plausible way of increasing the efficiency of this interaction might be to require would-be users of a system to demonstrate first a given level of typing skill. Secretaries who type business correspondence must, of course, have a high level of proficiency. On the other hand, proficiency in typing is required in some occupations where the need for it is much less apparent. For example, Federal Aviation Administration flight service specialists use keyboards only to teletype weather reports and flight plans. Yet they must demonstrate the ability to type at least 35 words a minute.

The usual method of evaluating typing performance is to measure a typist's keying speed and accuracy in copying text. Simple copying skills are, however, only a minor factor in most tasks involving communication via keyboard. Jurgensen (1942) reports that tests of simple copying skills have little validity in predicting typing success in stenographic and secretarial positions. By concentrating on the psychomotor aspects of key stroking such tests neglect a host of other important skills This is strong evidence for what West (1967 1969b) has referred to as the "substantia irrelevance of straight copy skills to proficiency at realistic typing tasks."

---

[1] This research was done under Research Grant GN35023 from the National Science Foundation, Office of Science Information Service.

[2] Requests for reprints should be addressed to Alphonse Chapanis, Department of Psychology, Johns Hopkins University, Baltimore, Maryland 21218.

In a realistic task involving interactive communication with a computer system, Morrill, Goodwin, and Smith (1968) found that subjects without typing skills performed their task about as fast as skilled typists and with no more errors. Unfortunately, their design confounded typing skill with sex and educational level. The typists were all female secretaries with high school educations, while the nontypists were males who had attended at least one year of college.

In a similar kind of study, Chapanis, Ochsman, Parrish, and Weeks (1972) had two-man teams of subjects communicate through typewriters to solve problems such as might be used in interactive computer systems. Some of the teams were made up of inexperienced typists, others of subjects who had completed a formal course in typewriting. While the linguistic transmission rate of the typists was greater than that of their untrained counterparts (18 versus 10 words per minute), the differences between the two groups were trivial in all other major respects—total solution time, behavioral activity, and verbal output. The investigators readily admit that this unexpected result may have been caused by unavoidable confounding in the execution of the experiment. Since the skilled typists were selected and tested later than the unskilled typists, the former may have come from a different population than the latter. In particular, no attempt was made to control for possible differences in verbal ability between the two groups. This may be a significant factor in view of the considerable evidence, beginning with the classic work of Bryan and Harter (1899), for a large verbal component in any text keying task.

The major purpose of this study was to provide an unconfounded test of the effect of differences in typing skill on dyadic communication via teletypewriter such as might characterize interactive man–computer dialogue. Another purpose was to collect verbal and behavioral data that might help explain why skilled typists do or do not enjoy an advantage over unskilled typists. The final purpose was to achieve much better control over the equivalence of the two groups of subjects than had characterized previous studies in the field. This equivalence was checked by having the skilled and unskilled typists also communicate in a mode simulating natural face-to-face conversation.

METHOD

Subjects

Sixteen female students, seniors at a parochial high school in Baltimore ranging in age from 17 years 5 months to 18 years 3 months, were hired as subjects. Subjects were paid a base rate of \$2.50 per session, a bonus of \$.50 for arriving at a correct solution, and, as an incentive to complete the four days of test, an additional bonus of \$3.00 on the last day of test.

The eight subjects in the "skilled" typist group were selected from secretarial courses and were certified by their instructor to have typing speeds in excess of 40 words per minute. Our eight unskilled typists were selected from the college-oriented curriculum of the school. Our pragmatic criterion was that the unskilled typists should have a typing speed of less than 25 words per minute. Although we recognize that it would have been desirable to set this cutoff score lower, in our society it is virtually impossible to find high school students who have never used a typewriter and yet who are comparable to skilled typists in all other respects.

To control for intellectual ability, all prospective subjects were given the verbal ability portion of the School and College Ability Test (SCAT), Form 2A, sections 1 and 3 (Educational Testing Service, 1957). The subjects were selected so that the skilled and unskilled typist groups were matched on this variable ($M = 300.2$, $SD = 10.6$; $M = 299.4$, $SD = 13.5$, respectively).

Apparatus

The two members of a team communicated from two adjacent, sound insulated rooms. The wall between the rooms contained three apertures. One was filled with a double-paned plate glass window (1.06 × 1.39 m.) especially constructed to be sound insulating. Speaker grill cloth was stretched across a second aperture (.75 × .20 m.) to allow virtually unrestricted voice communication. Finally, the third opening (.25 × .33 m.) was covered by two interconnected doors, designed so that opening the door in one room automatically closed the door in the other. This arrangement allowed objects to be passed between rooms but prevented direct visual contact through the passage. When all three apertures were left open, the subjects were able to converse face-to-face, a condition that we have been calling a communication-rich mode (Chapanis, 1973; Chapanis et al., 1972). When visually opaque, sound insulating panels were placed over the three apertures, the subjects were effectively isolated from each other.

Each room contained a work table and chair for the subject, and an IBM Model B input–output writer slaved to the corresponding machine in the other room. A message typed on one machine appeared simultaneously on the other. Each subject had a control box with which she could at any time interrupt her partner and take control of the typewriter communication channel or give control of it to her partner. A more detailed description of the experimental arrangements may be found in Chapanis et al. (1972).

Modes of Communication

Two different modes of communication, typewriting and communication-rich, were tested in this experiment. In the typewriting mode, subjects were permitted to communicate only through the slaved typewriters in each room. The window and grill cloth opening were covered so that they could neither see nor hear one another.

In the communication-rich mode, the subjects could both see and hear each other. They could talk freely; show each other their parts of the problem; and point, nod, grimace, or use any other form of verbal or nonverbal communication that they might use in normal face-to-face conversation. The only restriction was that the window and wall between them prevented a subject from touching her partner or her partner's problem-related materials.

Job Roles

The subjects on a team had two functionally different roles. The subject we refer to as the seeker was, in effect, the person with the problem to solve. Her information folio defined the nature of the problem, identified any constraints on it, and contained some of the information needed for its solution. The subject we refer to as the source served as an assistant to the seeker, providing the rest of the necessary information from her information folio. Thus, the cooperative efforts of both team members were needed to arrive at the solution.

Problem-Solving Tasks

Rather than use abstract, artificial problems of the kind so often exploited in problem-solving studies, we attempted to model ours after everyday, real-world situations. They were tasks on which two people might naturally work together, and for which the aid of present or future computer systems might be useful.

We selected four problems: (*a*) an object identification and (*b*) an information retrieval problem, both developed earlier by Chapanis and Overbey (1974), and (*c*) a faultfinding and (*d*) a class-scheduling problem, both developed especially for the present study.

The object identification problem. The seeker was given a small electric pilot light socket and was asked to obtain an identical replacement socket from her partner. The source was given a large tray containing 65 different Leecraft pilot light sockets arranged in individual receptacles. Although all 65 sockets in the source's folio were similar to the seeker's socket in some respect, only one was a perfect match.

The information retrieval problem. The seeker was asked to prepare a bibliography of five different newspaper articles relevant to the topic: The threat to individual privacy that would be posed by a United States Data Center based on a computerized information retrieval system. The source was given the New York Times Index for 1967, from which the citations were to be drawn.

The faultfinding problem. The seeker had to perform three diagnostic tests on a mockup of an automobile ignition system, using a commercially available automobile engine analyzer. She was told that there were one or more defects (there were actually two). The source was given a detailed set of instructions for the tests.

The class-scheduling problem. The seeker was given a list of four college courses which had to be arranged into a workable schedule within specified time constraints, such as commuting schedules. The source had a 97-page booklet listing the complete time schedule for courses at the University of Maryland. The courses and constraints were such that there was only a single correct solution.

Experimental Design

The design used two identical Graeco-Latin squares with skill levels as the between-squares factor (see Table 1). A between- and within-modes format was superimposed on one variable of the Graeco-Latin square. This design gave us a straightforward test of the two sources of variation in which we were most interested, the main effect of skill levels and the Skill Level × Mode interaction. We included the communication-rich mode as an additional check on the equivalence of our two subject groups. If our selection procedures were successful, the two groups of

TABLE 1
EXPERIMENTAL DESIGN

Skill levels	Teams[a]	Communication modes			
		Communication-rich		Typewriting	
Skilled typists	T_1	P_2[b], D_4[c]	P_4,D_2	P_1,D_3	P_3,D_1
	T_2	P_4,D_1	P_2,D_3	P_3,D_2	P_1,D_4
	T_3	P_1,D_2	P_3,D_4	P_2,D_1	P_4,D_3
	T_4	P_3,D_3	P_1,D_1	P_4,D_4	P_2,D_2
Unskilled typists	T_5	P_2,D_4	P_4,D_2	P_1,D_3	P_3,D_1
	T_6	P_4,D_1	P_2,D_3	P_3,D_2	P_1,D_4
	T_7	P_1,D_2	P_3,D_4	P_2,D_1	P_4,D_3
	T_8	P_3,D_3	P_1,D_1	P_4,D_4	P_2,D_2

[a] Each team, *T*, was made up of two girls, designated as seeker and source.

[b] P_1 (object identification), P_2 (class scheduling), P_3 (faultfinding), and P_4 (information retrieval) are the four problems.

[c] D_1, D_2, D_3, D_4 are the first, second, third, and fourth days of test.

TABLE 2

THREE FORMS OF THE ANALYSIS OF VARIANCE

Source of variation	df		
	A	B	C
Between teams (T)	7	7	7
Skill levels (SKL)	1	1	1
Between teams within skill levels (T/SKL)	6	6	6
Subjects within teams (S/T)		8	8
Job roles (JR)		1	1
JR × SKL		1	1
JR × T/SKL		6	6
Within subjects		48	16
Modes (M)	1	1	
M × T	7	7	
M × SKL	1	1	
M × T/SKL	6	6	
M × S/T		8	
M × JR		1	
M × JR × SKL		1	
M × JR × T/SKL		6	
Residual	16	32	
Total	31	63	31

subjects should solve problems about equally well in the communication-rich mode. In that case, the main effect of skill levels would be one test of any differential effect of typing skill. On the other hand, if our selection procedures did not succeed in perfectly matching the two groups of subjects, the lack of equivalence would be revealed by differential performance when conversing face-to-face. In that case, the effect of typing skill could still be evaluated through the Skill Level × Mode interaction.

Because modes were superimposed hierarchically on one variable of the Graeco-Latin squares, the test for modes may be confounded with a possible Problem × Day or Mode × Problem × Day interaction. However, we could see no reason to expect these interactions to occur. A previous study (Chapanis & Overbey, 1974) had found no interactions involving days. In any case, these two modes of communication have already been extensively compared in earlier studies (Chapanis et al., 1972; Chapanis, 1973; Ochsman & Chapanis, in press), so that they were not variables of primary interest here.

Three forms of analysis of variance used for the data are shown in Table 2. The rationale for the three different analyses is explained in the section on results and discussion.

Procedure

The eight subjects within each skill level were randomly paired to make four teams of skilled typists and four teams of unskilled typists. Each of the team members was then assigned one of the two job roles, that is, source or seeker. When they arrived for an experimental session, the team members remained together while the experimenters read to them the general instructions for the problem they were scheduled to solve that day. The instructions emphasized the importance of cooperation between the participants. The subjects were told to act naturally and to do or say anything that might expedite solution of the problem. They were to work as quickly as possible, and would be paid a bonus if their solution was correct.

The subjects were then separated and taken into the two experimental rooms. Each subject was given her problem-related materials and was familiarized with the communication apparatus. The problem began when the experimenters were satisfied that the subjects were familiar with the equipment, procedures, and task.

Data Collection

Three kinds of data were recorded: time to solution, observable behavior, and verbal output. All teams reached a solution to all problems. Since the few incorrect solutions (6 of 32) were distributed evenly between skill levels and among problems, incorrect solutions have been ignored in our analyses.

Behavioral measures. With minor revisions, our method of recording behavior was similar to that used by Chapanis et al. (1972). One experimenter recorded the behavior of the seeker, the other that of the source. Each experimenter heard an auditory tone in an earphone every five seconds. He then recorded what the subject had been doing at the instant of the signal.

The subject's behavior was recorded in one of nine categories. The activity of "sending only" was defined as engaging exclusively in the transmission of a message either by typing, by speaking, or by using some form of nonverbal communication. An activity was scored as "other sending" if the subject was simultaneously sending a message *and* engaging in some other activity. There were two analogous receiving activities—"receiving only" and "other receiving." The remaining five categories were: "searching only," "making notes," "parts handling," "waiting," and "other."

Two composite behavior categories were derived from the measures above. "All sending" was the sum of times spent in sending only and in other sending. "All receiving" was the sum of times spent in receiving only and in other receiving.

Following the procedure used by Chapanis et al. (1972), an independent observer sat in on one session with each of the two regular observers to check interobserver reliability. The obtained reliabilities (.87 and .96) are of the same magnitude as found in the earlier study.

Time to solution. The total time to solve a problem was calculated as the sum of the times a subject spent in the nine behavior categories, corrected for any observations missed by the experimenter.

Verbal output. Audio tape recordings were made of all sessions in the communication-rich mode. These recordings were transcribed and then verified against

TABLE 3

STATISTICALLY SIGNIFICANT SOURCES OF VARIATION IDENTIFIED BY THE 14 ANALYSES OF VARIANCE

Dependent variables	Sources of Variation						
	Skill levels (SKL)	Modes (M)	SKL×M	Job roles (JR)	SKL×JR	M×JR	SKL×M ×JR
Time to solution		.006		—	—	—	—
Behavioral measures							
Sending only		<.0005					
Other sending		—	—			—	—
Receiving only		.0005					
Other receiving		—	—			—	—
All sending		.045					
All receiving							
Searching		.06		.002			
Making notes	>.975			.001			
Waiting							
Verbal measures							
Number of messages		.035					
Number of words							
Message length		.0005					
Communication rate		<.0005		>.99			

Note. A dash denotes that the term is not a source of variation in the analysis of variance for that dependent variable.

the tape by two persons working independently: the secretary who typed the original transcript and one of the experimenters. Subjects in the typewriting mode generated their own hardcopy record of their communications. Several verbal measures (described later) were computed from these transcripts and records.

RESULTS AND DISCUSSION

All statistically significant effects obtained in this study are shown in Table 3. By far the most conspicuous finding in the table is the strong and consistent effect that modes exerted on almost every dependent variable. Our main interest, however, was in the main effect of skill levels and in the interaction of Skill Level × Mode. Neither of these two terms was significant for any dependent measure.

Faced with the statistical conundrum of having to prove the null hypothesis, we decided to show data for the two skill levels within each mode whenever we showed significant mode effects. While the figures and tables thus follow the conventional format for illustrating significant interactions, our purpose in presenting the data this way is exactly the opposite. Our aim is to show how unimportant typing skill is for interactive communication via typewriters.

Time to Solution

One important dependent measure in this study was the total time taken by a team to solve a problem. Because solution time is a measure of team performance and not individual performance, the analysis of variance of the time data (Form A, Table 2) contains no terms for job roles or any interactions involving job roles.

The height of each bar in Figure 1 represents the average time to solve problems for that skill level-mode combination. The most striking feature of that figure is the large difference between the two communication modes. On the average, our teams of subjects solved problems more than three and one-half times faster when conversing face-to-face than when communicating through the slaved typewriters. Incidentally, the difference in solution time between modes appears to be inflated by the extremely short times required to solve the object identification problem in the communication-rich mode ($M = 3.8$ minutes). If only the other three problems are considered, there is a 2.8:1 ratio between modes, a figure that is much closer to the 2.3:1 ratio Chapanis et al. (1972) found between their communication-rich and typewriting modes.

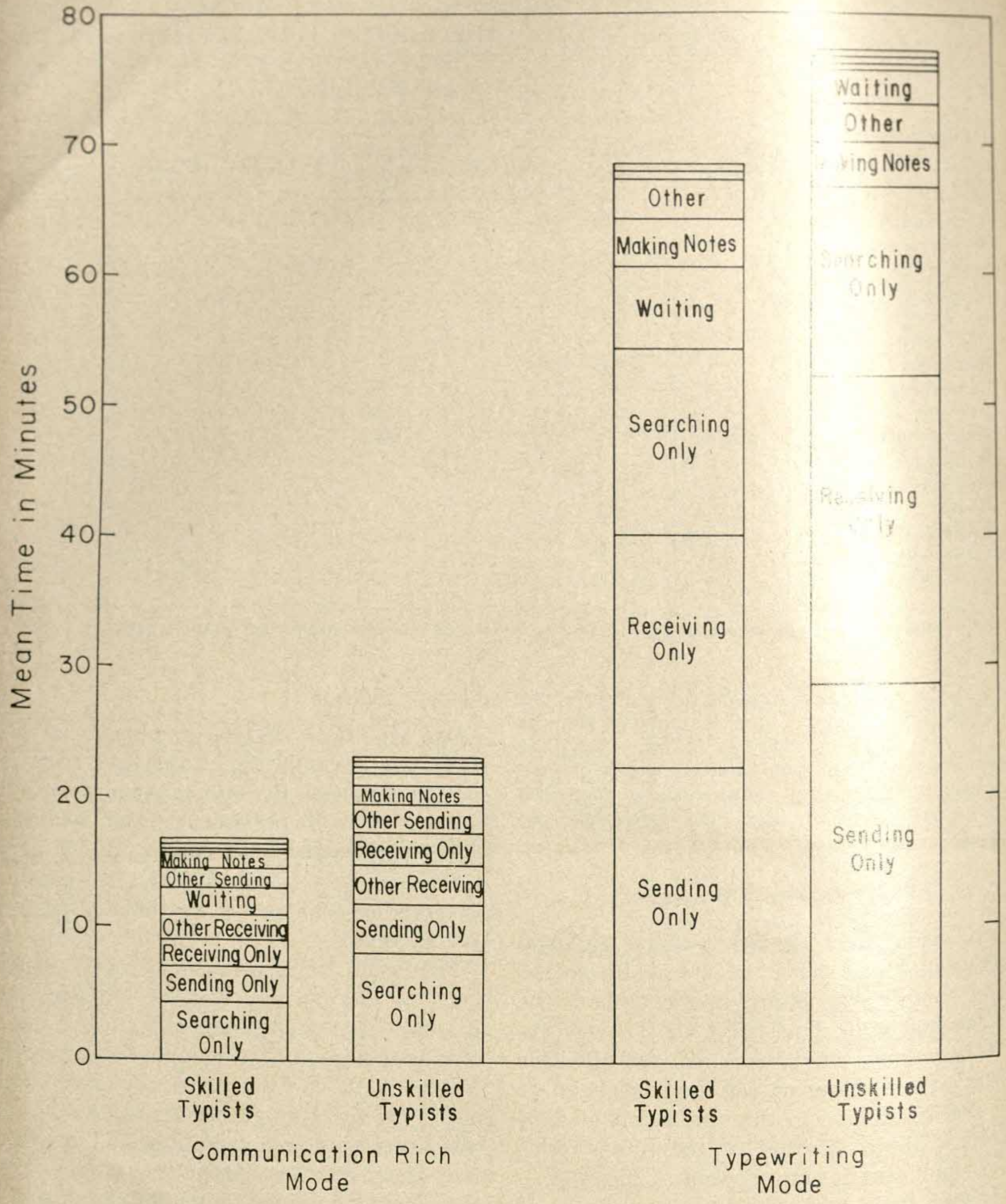


FIGURE 1. Mean time to solve problems for each skill level in the two communication modes with behavioral category time distributions.

Neither skill levels nor the interaction of Skill Level × Mode was significant. To be sure, the skilled typists solved problems about nine minutes faster than did the unskilled typists in the typewriting mode. However, the skilled typists were very nearly that much faster (about seven minutes) than their unskilled counterparts in the communication-rich mode as well. Although the skilled typists appear to have been able to solve these problems somewhat faster, there is no evidence whatsoever in these data that typing skill per se gave the skilled typists any differential advantage.

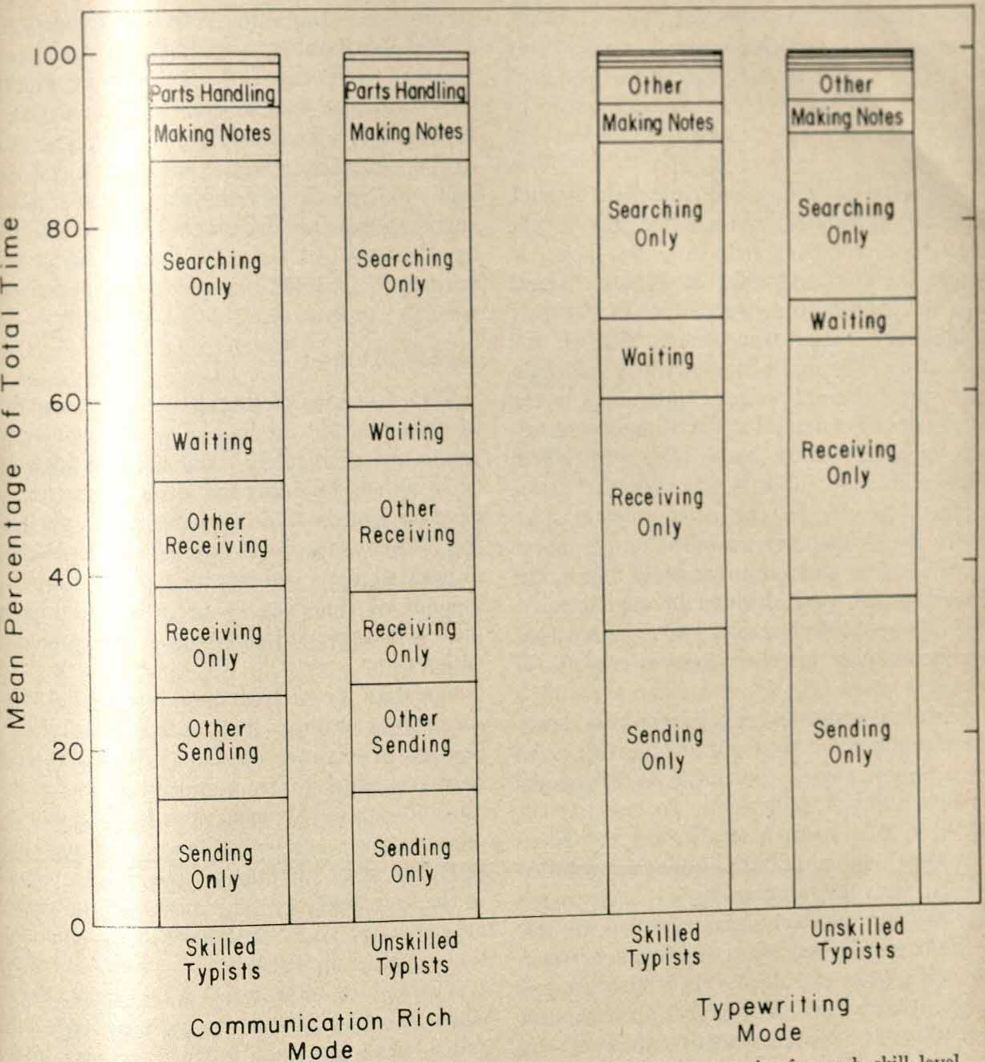


FIGURE 2. Mean percentage of total time spent in various behavioral categories for each skill level in the two communication modes. (The data of Figure 1 are replotted here as percentages.)

Behavioral Measures

Our activity analyses enabled us to see how the total times in Figure 1 were distributed among the nine behavioral categories. The segments of each bar in Figure 1 show, in rank order, the average times subjects spent in the various activities. Since the amount of time devoted to an activity was free to vary for each team member, each mean is based on a number of 16. Some categories occupied so little of the average subject's time that they are represented as unlabeled spaces at the top of each bar.

Because of the large differences between the two modes, the data were transformed to percentages to facilitate comparisons both within and between behaviors. To do this, the time devoted by each subject to a particular behavior was expressed as a percentage of the total time spent on a problem. Analyses of variance (Form B, Table 2) were performed on those percentage data for five of the nine individual behavioral categories and the two composite behavioral categories. Form C of the analysis of variance was used for other sending and other receiving because these two behaviors did not, by definition, occur in the typewriting mode. Parts handling and other were not analyzed since they each occupied too little of the average subject's time (less than 3%) to be of practical importance.

The mean percentages of time spent in each behavior, calculated in the same way as the mean times for each behavior, are shown in Figure 2. The categories in Figure 2 have been rearranged from their order in Figure 1 to highlight certain comparisons that we find interesting. The figure shows clearly that subjects spend their time quite differently in the two modes of communication. Equally impressive, however, is the consistency with which the two kinds of subjects apportion their time within a mode. In the communication-rich mode, the skilled and unskilled typists never differ by more than approximately 3% in the percentage of time devoted to any activity. This is one additional piece of evidence that we successfully matched the two groups of subjects.

The more important set of findings is shown in the two bars on the right in Figure 2. Even when typing, there is not much to distinguish between the two skill levels. To be sure, the skilled typists spent a smaller percentage of their time sending only than did the unskilled typists, but the difference, 3.4%, is much smaller than one might have expected in view of the large difference in typing skill between our two groups of subjects. It is also ironical to note that rather than exploit their modest advantage in communication efficiency by turning to other productive activities, the skilled typists offset their gain by spending a greater percentage of their time waiting (9.0% versus 4.5%) than did the unskilled typists. The sum of sending only plus waiting, in the typewriting mode, is 41.8% for skilled and 40.6% for unskilled typists. These sums replicate almost precisely those obtained by Chapanis et al. (1972) for their "experienced" (39.4%) and "inexperienced" (40.2%) typists.

TABLE 4

MEANS OF THE VERBAL MEASURES FOR THE FOUR SKILL LEVEL–COMMUNICATION MODE COMBINATIONS

Verbal measure	Communication-rich mode		Typewriting mode	
	Skilled typists	Unskilled typists	Skilled typists	Unskilled typists
Number of messages per subject	74	94	33	26
Number of words per subject	596	803	571	507
Message length (words per message)	8	9	17	19
Communication rate (words per minute of sending time)	122	147	27	18

Verbal Measures

Two measures of linguistic output, number of messages and number of words exchanged, were counted directly from the protocols for each subject. Dividing the latter by the former gave us a mean message length for each subject. Finally, the total number of words communicated by a subject was divided by the amount of time she spent in all sending to give us a second derived measure, communication rate.

The data for each of these four verbal measures were entered into a [illegible] B type of analysis of variance (Table 2). The difference between modes of communication was a significant source of variation for number of messages, message length, and communication rate. No other significant effects were found in the four analyses. The means of the verbal measures for each skill level-mode combination are given in Table 4.

Number of messages. A message began when a subject began to talk or type and ended when she had finished or was interrupted by her partner. Table 4 shows that subjects sent about 2.8 times as many messages when speaking as when typing. This ratio between modes becomes more dramatic when one recalls that the time devoted to message transmission in the communication-rich mode was only one-fifth that in the typewriting mode (Figure 1).

Number of words. Earlier studies in this research program had found that subjects in the communication-rich mode, or in any communication mode using a voice channel, have transmitted three to five times as many words as subjects in the typewriting mode (Chapanis, 1973; Chapanis & Overbey, 1974). The word counts in this experiment were in the same direction, but the difference between the modes is much smaller and not statistically significant.

Message length. Since there is a large and statistically significant difference between the number of messages communicated in the two modes, but only a small difference in the number of words used, it follows that message lengths must differ. The data in Table 4 confirm that subjects "packaged" their information differently in the two modes: Spoken messages were only about half as long as typewritten messages.

Communication rates. On the average, subjects communicated six times as many words per minute when they spoke as they did when they typed. This difference between modes is both statistically significant and large enough to be of practical importance, but not particularly surprising. The interesting feature of the data in Table 4 is the relatively small difference in communication rates between the two skill levels in the typewriting mode. On our pretest of typing skill, the skilled typists were far superior in performance to the unskilled typists (43 versus 21 words per minute). When typing to communicate information in the solution of problems, however, the skilled typists typed only nine more words per minute than did the unskilled typists.

Learning Effects

Table 5 presents the mean values of several dependent measures for the four days of test. Considering the ubiquity of learning effects in human experimentation, we expected that our teams would show some improvement in their ability to solve problems over four successive days of testing. Learning effects cannot, however, be assessed directly given the experimental design of this study. While problems and days are each balanced across the two communication modes, the unique problem-day combinations are not so balanced. As a result, the effect of days is confounded with a possible Mode × Problem interaction. Because of that confounding, we believe there is no defensible method of evaluating the statistical significance of the differences among the means shown in Table 5. Nonetheless, systematic changes in performance on any dependent measure are notably absent. There is some cross-validation for this observation in the study by Chapanis and Overbey (1974). With an experimental design that did permit a direct examination of the main effect of days, they found days to be significant on only one out of six dependent measures—message length.

TABLE 5

MEANS OF VARIOUS DEPENDENT MEASURES ON THE FOUR DAYS OF TEST

Dependent variable	Day of test			
	1	2	3	4
Time to solution (minutes)	51	47	54	34
Number of messages per subject	74	38	58	58
Number of words per subject	821	397	721	539
Message length (words per message)	14	11	14	14
Communication rate (words per minute of sending time)	77	83	77	78

Note. Each entry is the unweighted mean for 16 subjects.

Typing Skill

The apparently counterintuitive statement that typing skill does not significantly aid interactive communication via typewriters becomes less puzzling in light of two conclusions that may be drawn from the data. The behavioral data show that about one-third of the total problem-solving time is devoted to sending, that is, typing. Thus, any potential advantage in typing skill would be operating only one-third of the time. In addition, there was a marked drop in typing speed for the skilled typists from pretest ($M = 43$ words per minute) to experimental performance ($M = 27$ words per minute). The corresponding change for the unskilled typists was much less, that is, from 21 to 18 words per minute.

The sharp decline in typing speed for the skilled typists may be attributed to the nature of the task, or, more specifically, the nature of the material typed. Muhich (cited in West, 1969a) found that second-semester high school typists with an average speed of 41.7 words per minute on straight copy work typed at less than half that speed when they typed material such as letters, tables, and rough drafts. West (1969a), in discussing Muhich's (cited in West, 1969a) data, attributed the decline in speed to a different set for production-task stroking behavior and to the increased role of planning and decision-making skills in working with material that must be organized as well as typed. Our pretest of typing skill, intentionally like those customarily used to select typists, was a measure of speed and accuracy in copying prepared text. By contrast, our experimental task, while not like the production tasks used by Muhich (cited in West, 1969a), required a great deal of planning and decision making as our subjects composed their messages at the typewriter.

CONCLUSION

In summary, it appears that the efficiency of typewriter communication for interactive problem solving is determined not so much by how fast one can type as by how quickly one can formulate what to type. One cannot, of course, prove the null hypothesis. However, in the light of our results and those of studies cited earlier, we conclude that a typing skill prerequisite for using typewriter terminals would produce, at best, only modest gains in systems efficiency.

REFERENCES

Bryan, W. L., & Harter, N. Studies on the telegraphic language: The acquisition of a hierarchy of habits. *Psychological Review,* 1899, **6**, 345–375.

Chapanis, A. The communication of factual information through various channels. *Information Storage and Retrieval,* 1973, **9**, 215–[illegible].

Chapanis, A., Ochsman, R. B., Parrish, R. N., & Weeks, G. D. Studies in interactive communication: I. The effects of four communication modes on the behavior of teams during cooperative problem solving. *Human Factors,* 1972, **14**, 487–509.

Chapanis, A., & Overbey, C. M. Studies in interactive communication: III. The effects of similar and dissimilar communication channels and two interrupt options on team problem solving. *Perceptual and Motor Skills, Monograph Supplement* 2-V38, 1974, **38**, 343–374.

Educational Testing Service. *Cooperative school and college ability tests.* Princeton, N.J.: Educational Testing Service, 1957.

Irwin, M. R. The computer utility. *Datamation,* 1966, **12**(11), 22–27.

Jurgensen, C. E. A test for selecting and training industrial typists. *Educational and Psychological Measurement,* 1942, **2**, 409–425.

Michaels, S. E. QWERTY versus alphabetic keyboards as a function of typing skill. *Human Factors,* 1971, **13**, 419–426.

Morrill, C. S., Goodwin, N. C., & Smith, S. L. User input mode and computer aided instruction. *Human Factors,* 1968, **10**, 225–232.

Ochsman, R. B., & Chapanis, A. The effects of ten communication modes on the behavior of teams during cooperative problem solving. *International Journal of Man-Machine Studies,* in press.

West, L. J. Production proficiency among typists: Research and implications. *Business Education Forum,* 1967, **22**(2), 5–7.

West, L. J. *Acquisition of typewriting skills: Methods and research in teaching typewriting.* New York: Pitman, 1969. (a)

West, L. J. Fatigue and performance variability among typists. *Journal of Applied Psychology,* 1969, **53**, 80–86. (b)

(Received February 11, 1974)

Journal of Applied Psychology
1974, Vol. 59, No. 6, 675–678

EFFECTS OF GROUPING OF PRINTED DIGITS ON FORCED-PACED MANUAL ENTRY PERFORMANCE[1]

E. T. KLEMMER[2] AND L. P. STOCKER[3]

Bell Laboratories, Holmdel, New Jersey

Previous studies have shown that optimum grouping of printed digits leads to increased speed in keyboard entry, but differences in error rate due to grouping have generally failed to reach statistical significance. The present study presented numbers to be keyed at a forced-paced rate, chosen separately for each subject from his self-paced rate so as to provide moderate speed stress. Under this condition ($n = 18$), the error rate was doubled by using a 1–5–1 grouping of seven-digit numbers as opposed to a 3–4 grouping. For five-digit numbers, which cannot be grouped by 3s and 4s, no significant difference in performance was found ($n = 24$) between 3–2 grouping and ungrouped presentation.

Several previous studies have shown that the highest speed in keying visually presented numbers occurs when the digits are grouped by 3s and 4s (Klemmer, 1969). Differences in error rate due to temporal grouping of auditory digits can be very large, as shown by Ryan (1969), who employed spoken recall, but studies which have used visual stimuli in self-paced tasks have shown little difference in error rate as a function of grouping. Since some keyboard devices (e.g., mail-sorting machines) operate at fixed speeds, it is important to ask what happens to error rate as a function of grouping when a fixed presentation rate is employed with visual stimuli.

The present experiments examined the effect of grouping upon error rate under forced pacing in which rate of presentation is constant for all trials and groupings. The constant forced rate used in the present study was different for each subject (set to match his particular skill level on the basis of preceding self-paced trials). Two separate experiments were conducted, one with seven-digit numbers and another with five-digit numbers. Seven-digit numbers were used because of the large amount of self-paced data previously gathered with sequences of this length. Five-digit numbers are of interest because they cannot be grouped in 3s and 4s and are in common use, notably as zip codes.

METHOD

Apparatus

Numbers were presented one at a time on the screen of a Model IIB Picturephone® set. This was done by having the camera of a sending Picturephone set focused through a 132-millimeter lens on the printed output of a Model KSR35 teletypewriter which was in turn controlled by a Tempo computer. The subject responded by pressing keys on a Touch-Tone® keyboard which was also connected to the computer for automatic timing and scoring of accuracy. Figure 1 shows the subject's position with display Picturephone set and Touch-Tone keyboard.

Stimuli

Seven-digit numbers were presented in Experiment 1 with one of three groupings chosen on the basis of past studies to represent the best (3–4), worst (1–5–1), and intermediate (no grouping) for seven-digit numbers in terms of speed of entry. Typical numbers appeared as follows:

736 2311
3 88912 4
6723985

Five-digit numbers were presented in Experiment 2 either ungrouped or grouped 3–2, as was done by Brown and Cornog (1970). Typical numbers appeared as shown:

43520
149 22

The presented digits were randomly chosen without restriction. The size of the number on the screen was such that the horizontal extent ranged from 6.7 centi-

---

[1] Experiment 1 of the present study was presented at the 80th Annual Convention of the American Psychological Association, Honolulu, September 4, 1972.

[2] Requests for reprints should be sent to E. T. Klemmer, Bell Laboratories, Holmdel, New Jersey 07733.

[3] Now at the Chesapeake and Potomac Telephone Company, Washington, D.C.

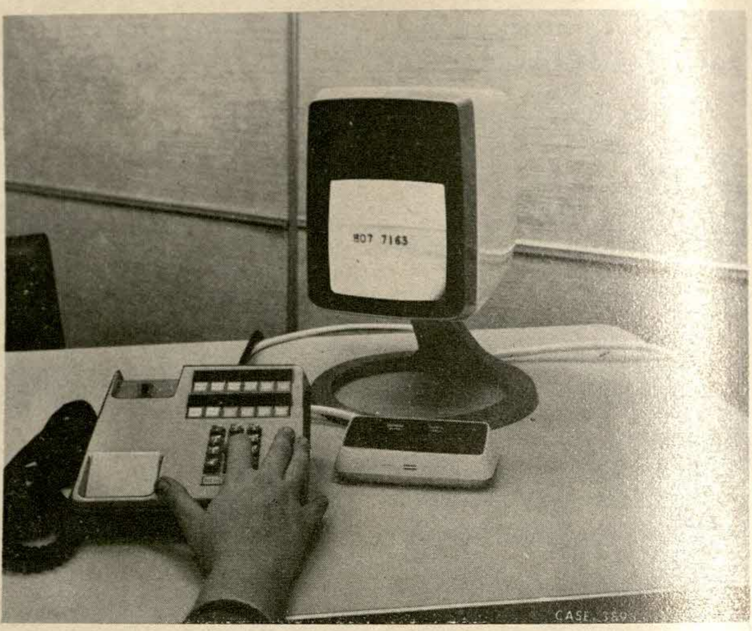


FIGURE 1. View of display and keyboard from subject's position.

meters for the 1–5–1 grouping to 3.6 centimeters for the five-digit ungrouped numbers. Viewing distance was about 50 centimeters. No control of the position of the subject's head was exercised.

Subjects

Eighteen employees of Bell Laboratories served as subjects in Experiment 1 (with seven-digit numbers): Six were randomly selected engineers, 6 did clerical work, and 6 were shop workers. The clerical subjects were selected by supervisors from clerical and typing pools, and the shop subjects were selected by supervisors from machine and wiring shops. Twenty-four subjects served in Experiment 2 (with five-digit numbers), 18 of whom were employees with the same job classifications as those in Experiment 1, plus 6 housewives.

Timing and Scoring

During the self-paced runs, a new number would appear .2 seconds after the final digit was keyed without regard to accuracy. For the forced-paced condition, the numbers changed at a constant time interval set by the experimenter at the start of the run. The computer compared the keyed digits with those presented and printed out all errors at the end of the run for both the self-paced and forced-paced conditions.

The scoring interval in the forced-paced condition was set to end 100 milliseconds after the number disappeared. In addition, a short tone (150 milliseconds) was sounded just before the disappearance of each number to warn the subject of the end of the scoring interval. The results showed this to be a satisfactory procedure to avoid out-of-phase responses.

The computer calculated a mean and variance for the keying time for all correctly entered numbers in a self-paced run. These times were used to compute the timing for the forced-paced runs.

Procedure

All subjects were first verbally given the self-paced instructions, which were as follows:

> In all trials, a number will appear on this Picturephone screen. You are to key the number as quickly and accurately as possible using this Touch-Tone telephone. As soon as you hit the last digit of the number, it will disappear and a new number will *immediately* flip into view, although you may not see it change if you are looking at the keyboard. Begin keying the new number immediately. Do not pause or attempt to correct previous errors. If you realize you made an error, there is no way to correct it, so just press the number sign key[4] [demonstration] to make the next number appear. Press the number sign key if you finish keying and the number does not change. Each of the blocks of trials will contain 50 numbers [40 in Experiment 2]. Begin keying as soon as the first number appears.

In Experiment 1 subjects were given 10 self-paced warm-up trials, grouped 2–5, followed by three separate runs of 50 trials each, with one run at each grouping. In Experiment 2 warm-up consisted of 5 self-paced trials, ungrouped, followed by four runs of 40 trials each, two of which were grouped and two ungrouped. The order was counterbalanced over subjects.

Next, the instructions for the forced-paced runs were given to the subject.

---

[4] The number sign key is the 12th key on the lower right-hand corner of new Touch-Tone telephones.

> On the basis of your self-paced keying speed, a keying time has been calculated. This will be used as the rate at which the numbers on the screen change. In other words, they will be regularly changing at a pace which is independent of when you now key in the number. Just before the number will change, there will be a short tone sounded from the phone. Stop keying that number even if you have not finished because otherwise they will be counted toward the next number.

Each subject then had forced-paced runs, one at each grouping in the reverse order of the self-paced runs.

For each subject all forced-paced runs were conducted at a single speed or timing determined as follows: The mean time between numbers plus 1.28 standard deviations was computed for each of the self-paced runs. The minimum of these values was then used as the constant time between stimuli in all forced-paced runs. If we assume a normal distribution of times within each self-paced run, this formula gives a time which is exceeded by 10% of the response times for that run. If the subject maintained this assumed normal distribution of response times in the forced-paced run, then a 10% error rate would be expected based on slow responses alone for the best grouping and even more errors for the other groupings. Actually, of course, the subject is capable of reducing the number of slow responses during the later forced-paced condition, and so increases in error rate by 10% are not reasonable predictions.

At the end of each subject's session, which lasted an average of 40 minutes, he was asked which grouping he preferred.

RESULTS

Figure 2 shows the mean time and error rate for each grouping under the self-paced condition and the mean error rate for each grouping under the forced-paced condition of Experiment 1. The differences in speed with grouping agree with the previous data in showing 3–4 best, ungrouped intermediate, and 1–5–1 poorest with .75 second, or 13%, increase in entry time compared to 3–4. All differences in speed between groupings were statistically significant.

The 1–5–1 condition also showed the highest error rate of the three groupings under the self-paced condition. This difference between 1–5–1 and 3–4 was statistically significant; the differences between each of those groupings and the ungrouped condition were not.

The error rate differences between groupings were larger under the forced-paced condition. The 1–5–1 grouping resulted in twice as many errors, on the average, as the 3–4 grouping. This result was not due to a few subjects, since 16 of the 18 subjects made more errors with 1–5–1 grouping than with 3–4 under forced pacing. The ungrouped condition again fell in between 3–4 and 1–5–1. All differences in error rates under forced pacing were statistically significant.

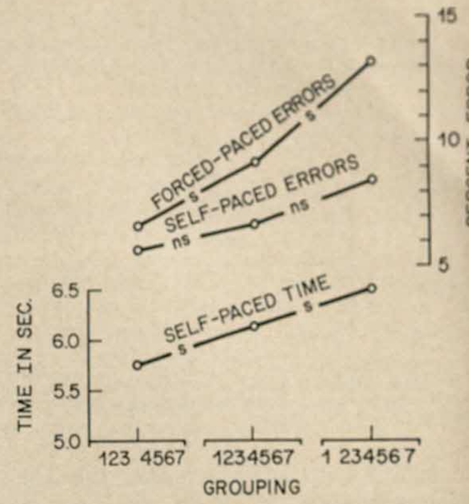


FIGURE 2. Keying time and accuracy as a function of grouping. (Each point represents the average data from 18 subjects, each of whom keyed 50 seven-digit numbers. The s or ns between two points indicates whether or not the difference between those points is significant at the .05 level. Time and errors are calculated per seven-digit number.)

All 18 subjects said they preferred the 3–4 grouping to the 1–5–1, and all but one preferred 3–4 to the ungrouped. The subjects divided evenly in preference between the 1–5–1 and ungrouped conditions.

Comparing the self-paced and forced-paced conditions for the 3–4 grouping (the forced-paced speed was set on the basis of the 3–4 self-paced speed in each case), we see that there was an increase in error rate from 5.6% for self-paced to 6.6% for forced-paced conditions over all subjects. This is much less than the 10% increase which would be predicted by assuming that the distribution of keying times was the same in both tests. Indeed, only 8 of the 18 subjects showed any increase in errors between 3–4 self-paced and 3–4 forced-paced conditions, so the overall difference was not statistically meaningful for the total population of subjects. These results suggest that people can eliminate some of their slow responses when forced, although

TABLE 1

TIME AND ERRORS IN KEYING FIVE-DIGIT NUMBERS AS A FUNCTION OF GROUPING UNDER SELF-PACED AND FORCED-PACED CONDITIONS

Condition	Self-paced time (in seconds)	Forced-paced time (in seconds)	Self-paced errors (in %)	Forced-paced errors (in %)
Grouped 3–2	3.76	4.32	4.3	4.1
Ungrouped	3.81		4.7	4.7

Note. Table entries are means of 24 subjects each of whom keyed 80 numbers under each condition. Time and errors are calculated per five-digit number.

confounding with practice makes this conclusion tentative.

Table 1 shows the mean times and error rates for the five-digit numbers of Experiment 2. The overall results for the self-paced trials showed small and statistically insignificant differences between the grouped and ungrouped conditions. Each subject had the two conditions in counterbalanced order, so it is meaningful to consider differences between groupings for each subject. Of the 24 subjects, 15 keyed faster with the 3–2 grouping, and 9 faster with the ungrouped numbers (a statistically insignificant difference). Error comparisons were made for both the self-paced and forced-paced test, 11 subjects made less errors with the 3–2 grouping (9 made more and 4, an equal number). Again these differences were statistically insignificant. Thus, the results generally favorable to the 3–2 grouping are all small and might well be due to chance.

These results are different from those of Brown and Cornog (1970), who found fewer overall errors for the ungrouped numbers. The difference may be due to the fact that they used short tachistoscopic exposures, whereas in the present experiment the stimulus was available for several seconds.

When asked which grouping they preferred, 20 of the 24 subjects stated a preference for the 3–2 grouping. This clearly significant result is surprising in view of the small performance differences between groupings.

DISCUSSION

It is clear that under conditions of forced pacing, the effect of grouping upon errors becomes larger and more consistent than under the self-paced condition. Since any printed digits are likely to be entered under forced speed by some people some of the time, the clear suggestion is that optimum (3s and 4s) grouping should be used for all printed digits, even if time saving under self-paced conditions is not an important factor for particular applications. A casual survey of the way digits are grouped in document numbers, credit cards, account numbers, part numbers, etc., shows that this simple rule is often overlooked. For five-digit numbers which cannot be grouped by 3s and 4s, it seems to make little difference in performance whether they are ungrouped or grouped 3–2, although subjects prefer the 3–2 grouping. The small performance differences should not be taken to mean that long sequences of digits can be grouped by 5s, however. Previous work has shown that it is not desirable (Klemmer, 1969).

Forced pacing in the present study was done by adjusting the pacing speed according to each subject. In many operating systems such individual adjustment is not possible, but rather all stations must work as the same speed. A reasonable prediction to be made from the present data for the single-fixed-speed case is that a greater percentage of people would fail to operate satisfactorily if less than optimum groupings were used. That is, for any criterion level of errors, more people meet the criterion with optimally grouped stimuli.

Subjects preferred the 3–4 grouping as well as performed better with it, but preferences did not distinguish between 1–5–1 and seven digits ungrouped. Subjects preferred the 3–2 grouping over five digits ungrouped, but performance differences were small.

REFERENCES

Brown, H. N., & Cornog, J. R. *Perception rates for zip codes as measured by performance on a tachistoscope.* (National Bureau of Standards Rep. 10 281) Washington, D.C.: U.S. Government Printing Office, 1970.

Klemmer, E. T. Grouping of printed digits for manual entry. *Human Factors,* 1969, 11, 397–400.

Ryan, J. Grouping and short-term memory: Different means and patterns of grouping. *Quarterly Journal of Experimental Psychology,* 1969, 21, 137–147.

(Received October 1, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 6, 679–682

EFFECTS OF READABILITY OF CONTEXT PASSAGES ON SUBSEQUENT INSPECTION RATES

ERNST Z. ROTHKOPF[1] AND RICHARD P. COATNEY[2]

Bell Laboratories, Murray Hill, New Jersey

College students ($N = 92$) studied a 1,200-word (19-paragraph) passage prepared in a standard (Flesch Reading Ease, $RE = 51$) and a difficult ($RE = 20$) version. Half of the subjects switched after the first 10 paragraphs from the standard to the difficult version or vice versa. The remainder saw all 19 paragraphs in the same (standard or difficult) form. Inspection rates were found to depend not only on the readability of currently inspected text but also the readability of recently inspected text segments. The last 9 paragraphs in standard form were inspected more slowly if the first 10 paragraphs were difficult than if the passage had been in the standard version throughout. Smaller context effects on inspection rate were observed when subjects were switched from standard to difficult material.

The relationship between the readability of text material and reading rate is both theoretically and practically important. Theoretical interest stems from attempts to understand the decoding process used by individuals in reading, while practical concern derives from efforts to measure and enhance the effectiveness of written communications.

Most previous experimental studies (see Klare, 1974) have operated under the tacit assumption that, outside of situational factors such as motivation, reading (inspection) rate at any given time is primarily determined by the difficulty of the text elements which currently confront the reader. The present experiment explores the possibility that effects of the readability of a given text element on inspection rate[3] persist over time. Specifically, the experimental question was whether the reading ease (Flesch, 1949) of recently inspected text influences the rate at which a subsequently presented text element is read.

This question is of some practical concern because any substantial context effects produced by readability would focus attention on the variability of reading ease in text materials. At present, the average readability index of a particular text material is the only characteristic typically considered in measuring the readability. However, Coke and Rothkopf (1970) have observed a wide range of readability values in 200 samples drawn from a single chapter of a science text.

The experimental exploration of "setlike" effects produced by text style is of general interest. For example, such studies may help to shed some light on discrepant results from studies of the relationship between readability and reading rate. Many of the early experiments in this area, which reported decreases in reading rate with lowered readability, must be disregarded because they used "words per minute" as a measure of rate. This measure is not suitable because average word length (in syllables or letters) is greater in the more difficult passages. Experiments that employed rate measures which were not affected by the dependency between readability and word length (e.g., syllables or letter spaces per unit time) have not produced consistent results. Three studies have found that inspection rate was relatively constant over a broad range of readability levels (Coke, 1974; Miller & Coleman, 1971; Sticht, 1971), and one (Carver, 1971–72) has reported mixed results. The exception was the result reported by Kershner (1964), who found that greater readability resulted in faster inspection speeds as measured in the number of type spaces per unit of time.

---

[1] Requests for reprints should be sent to Ernst Z. Rothkopf, Learning and Instruction Research Department, Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974.

[2] Now at Ohio State University.

[3] The term *inspection rate* is used in preference to *reading rate* throughout this article because of the ambiguous character of the term *reading*. Inspection rate is more neutral with respect to demands made upon subjects by the experimental situations.

All of these studies involved comparisons within subjects over varied text materials. Coke (1974) used 32–90 passages per subject, Miller and Coleman (1971) used 36, Sticht (1971) used 11, and Carver (1971–72) 8 passages of widely different readability levels per subject. By contrast, Kershner's (1964) subjects read only *2* passages in quick succession before being interrupted by a comprehension test, and the total number of passages per subject was limited to 4.

The failure to observe influences of readability on inspection rate in experiments in which the subject was asked to read a large number of passages could have been due to persistent effects of readability of prior parts of the material read. If many short passages of widely different readability levels for each subject were used, any persistence in the effects of readability on inspection rate may have masked inspection effects produced by the individual passages; that is, the average readability of all recently read text elements may have contributed to current inspection rate.

The purpose of the present experiment was to test the conjecture that inspection rate has contextual determinants. The general method was to measure whether the inspection rate of a given block of text was influenced by the reading ease of preceding text materials.

METHOD

Materials

The experimental material was derived from a 1,271-word passage on the geography and history of Thailand (U.S. Department of Defense, 1965). The passage consisted of 19 paragraphs. Two versions of the passage, standard and difficult, were used. The standard passage was substantially identical to that used in a previous experiment (Rothkopf & Coke, 1968). Flesch Reading Ease, as estimated by computerized procedures (see Coke & Rothkopf, 1970), was on the average, 51.03 per paragraph ($SD = 10.67$). The difficult version was prepared by rewriting the standard passages so as to use more complicated sentence constructions and rarer word forms but maintain the same paragraph divisions and approximately the same paragraph lengths in words. The average Flesch Reading Ease score per paragraph for the difficult passage was 20.14 ($SD = 14.38$).

Design

The experiment was run in two parts. In Experiment 1, half of the subjects ($n = 23$) saw all 19 paragraphs in the standard version (standard–standard treatment). The remaining subjects ($n = 23$) saw the first 10 paragraphs in the difficult version and the final 9 in the standard condition (difficult–standard treatment). In Experiment 2, half of the subjects ($n = 23$) saw all 19 paragraphs in the difficult version (difficult–difficult); the remaining subjects ($n = 23$) saw the first 10 paragraphs in the standard condition and the final 9 in the difficult version (standard–difficult).

After the entire text had been inspected, a completion test (Cloze) was administered consisting of the 19 paragraphs, each in the version seen during the experiment, with nine substantive (i.e., important nouns, adjectives, and verbs) terms deleted from each paragraph.

Apparatus and Procedure

A negative 35-millimeter lantern slide was prepared from each paragraph. A Kodak Carousel projector was used to project these paragraphs from the rear onto a 41.9 × 30.5 centimeter plastic-embedded screen (Polacoat LS75PL) that stood on a table approximately 60 centimeters in front of the seated subject.

The onset and offset of each slide was controlled by each subject with a hand switch that was connected to the Carousel projector and an electromechanical shutter system. A Hewlett-Packard 5233L electronic counter complete with Model 562A digital recorder automatically recorded the time each slide was inspected.

The subjects were directed to read each text slide as rapidly as they could and still learn the essential elements of the content. The pencil-and-paper completion test was administered in a room adjacent to the room in which the text slides were viewed.

Subjects

Paid volunteer students ($N = 92$) from Fairleigh Dickinson University in Madison, New Jersey, were assigned in equal number to the four conditions used in Experiments 1 and 2.

RESULTS

Experiment 1

The results from Experiments 1 and 2 were analyzed separately because they were not performed at the same time and because they involved somewhat different populations of subjects.

Inspection speed in syllables per minute was calculated separately for each paragraph. In Figure 1 the results of Experiment 1 are plotted for the standard–standard and difficult–standard treatments.

The greater reading ease of the materials seen in the standard–standard treatment in

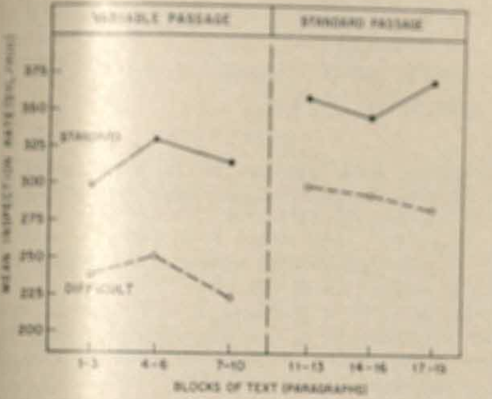


FIGURE 1. Mean inspection rate in syllables per minute for subjects exposed to standard and to difficult materials in the first 10 paragraphs of the text. (The last 9 paragraphs were at the standard readability level. The solid-line data correspond to the standard–standard and the dotted-line data to the difficult–standard treatment in the text.)

the first half of the text resulted in inspection speeds that were approximately 75 syllables per minute greater than those of the difficult–standard condition for paragraphs 1–10.

The observations that were most critical for the context hypothesis were the rates at which paragraphs 11–19 were inspected. These paragraphs were the same for both treatments and at the standard readability level. The results were consistent with the context hypothesis in that the differences between treatments in inspection rates in paragraphs 1–10 tended to persist on paragraphs 11–19. Average inspection rate for paragraphs 11–19 was 292.4 syllables per minute for the difficult–standard and 358.9 syllables per minute for the standard–standard treatment. The difference between the two treatments tended to maintain itself throughout the second half of the text.

The above conclusions were supported by results of an analysis of variance of average inspection speeds for each subject on paragraphs 11–13 and 17–19. The paragraph blocks at the beginning and the end of the critical section of the passage were used in this analysis in order to test whether reliable context effects persisted until the end of the text. Treatment effects were significant ($F = 5.09$, $df = 1/44$, $p < .05$), but the effects due to the paragraph set (i.e., paragraphs 11–13 versus 17–19) and the interaction between paragraph effect and treatment were not statistically reliable.

It should also be noted that the average inspection rate for paragraphs 11–19 in the standard–standard condition was about 44 syllables per minute greater than for paragraphs 1–10 even though the two blocks of paragraphs were at the same readability level for this treatment. This probably reflects the slow acceleration in inspection time during prolonged reading which has been observed in previous experiments (e.g., Rothkopf & Bisbicos, 1967).

The slower inspection speeds on paragraphs 11–19 in the difficult–standard treatments were accompanied by higher performance on those paragraphs on the completion test. Mean number of correct responses per paragraph for paragraphs 11–19 on the completion test were 4.08 for the difficult–standard treatment and 3.54 for the standard–standard condition. Mean correct responses for paragraphs 1–10 was 2.43 for the difficult–standard and 2.93 for the standard–standard condition.

A similar analysis of variance was carried out on the completion test data as on inspection speed, with the correct responses for paragraph sets 11–13 and 17–19 tallied separately. The analyses support the conclusion that subjects with the difficult–standard treatment learned more than the standard–standard subjects. Treatment effects were significant ($F = 6.74$, $df = 1/44$, $p < .05$). Neither the effect of paragraph sets nor the Paragraph Set × Treatment interaction was statistically reliable.

Experiment 2

The subjects who inspected the more readable material in the first half of the text (standard–difficult), inspected paragraphs 11–19 (difficult reading level) somewhat more rapidly ($\bar{X} = 291.5$ syllables per minute) than those in the difficult–difficult condition ($\bar{X} = 269.9$ syllables per minute). However, the difference between treatments, although consistent with the context hypothesis, was not statistically reliable (treatment $F = 2.97$, $df = 1/44$, $.10 > p > .05$). Average inspection rate for paragraphs 1–10 was 305.5 syllables per minute for the standard–difficult and 248.3 syllables per minute for the difficult–difficult treatment.

The data from the completion test indicated that the number of correct responses of paragraphs 11–19 were almost identical for the standard–difficult ($\bar{X} = 2.13$) and the difficult–difficult treatments ($\bar{X} = 2.17$). Mean correct responses for paragraphs 1–10 was 2.93 for the standard–difficult and 2.28 for the difficult–difficult treatments.

DISCUSSION

The results from Experiment 1 provide substantial indications that the rate at which subjects inspect text depends not only on the reading ease of the materials with which they are currently confronted but also the reading ease of recently inspected text segments. The results from Experiment 2 suggest that the context effect is probably stronger when subjects follow difficult reading with easier materials than when the switch is from medium to difficult readability.

The slower inspection rates induced by previous exposure to difficult materials were accompanied by relatively small but significant increases in learning. However, this effect cannot be unequivocally attributed to modification of inspection activity by context, since the possibility cannot be ruled out that the difficult–standard treatment subjects' performance on the test was influenced by learning level attained on the first 10 paragraphs when readability was different for the compared treatments.

The present experiment indicates fairly clearly that investigations of the relationship between readability and inspection speed must be interpreted with caution if the experiments employ short passages of widely different readability levels.

REFERENCES

Carver, R. P. Evidence for the invalidity of the Miller-Coleman readability scale, *Journal of Reading Behavior*, 1971–72, **4**, 42–47.

Coke, E. U. The effects of readability on oral and silent reading rates. *Journal of Educational Psychology*, 1974, **66**, 406–409.

Coke, E. U., & Rothkopf, E. Z. Note on a simple algorithm for a computer-produced reading ease score. *Journal of Applied Psychology*, 1970, **54**, 208–210.

Flesch, R. *The art of readable writing*. New York: Harper, 1949.

Kershner, A. M. Speed of reading in an adult population under differential conditions. *Journal of Applied Psychology*, 1964, **48**, 25–28.

Klare, G. R. Some empirical predictors of readability. In E. Z. Rothkopf & P. E. Johnson (Eds.), *Verbal learning research and the technology of written instruction*. New York: Teachers College Press, 1971.

Miller, G. R., & Coleman, E. B. The measurement of reading speed and the obligation to generalize to a population of reading materials. *Journal of Reading Behavior*, 1971–72, **4**, 48–56.

Rothkopf, E. Z., & Bisbicos, E. E. Selective facilitative effects of interspersed questions on learning from written materials. *Journal of Educational Psychology*, 1967, **58**, 56–61.

Rothkopf, E. Z., & Coke, E. U. Learning about added sentence fragments following repeated inspection of written discourse. *Journal of Experimental Psychology*, 1968, **78**, 191–199.

Sticht, T. G. *Learning by listening in relation to aptitude, reading, and rate-controlled speech: Additional studies.* (HumRRO Tech. Rep. 71-5). Alexandria, Va.: Human Resources Research Organization, April 1971.

United States Department of Defense. *A pocket guide to Thailand*. Washington, D.C.: U.S. Government Printing Office, 1965.

(Received December 26, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 6, 683–686

FORMATS FOR PRESENTING PROCEDURAL INSTRUCTIONS

ARTHUR S. BLAIWES[1]

Naval Training Equipment Center, Orlando, Florida

Thirty male college students performed a mock communication controller's task under three different instructional formats (short sentences, logical tree, and coding). Accuracy did not differ as a function of instructional format for three "easier" types of problems. However, for two "more difficult" problem types, coding resulted in significantly ($p < .01$) more accuracy in one case, while coding and short sentences showed significantly ($p < .01$ and $p < .05$, respectively) greater accuracy in another. It was concluded that (*a*) format variations mainly influence difficult tasks; (*b*) logical-tree superiority appears to be related to the number of conditions at a choice point and/or retention conditions; and (*c*) coding can be a potent technique.

The present experiment was concerned with improving the quality of printed instruction by influencing global characteristics of the presentation format. For example, instructions can be presented as continuous prose, an outline, flow diagrams, tables, etc. Relevant research has been restricted to comparisons mainly between only two presentation formats—logical trees (sometimes called algorithms, flow diagrams, or decision trees) and prose (Davies, 1970; Jones, 1968). Although logical trees typically prove superior to prose, some quite fundamental questions remain unanswered. One very basic question asks whether logical-tree superiority is due to the abstruseness of the prose employed (in which case any number of alternate methods for formatting the information would do as well).

The present experiment was designed to shed some light on such questions by comparing immediate and retention performance resulting from the use of logical trees to that resulting from using a version of a prose format, lists of short sentences. In a third condition, an effort was made to improve upon previous designs of logical trees by supplying short names ("codes") for the task rules, which subjects associated with the rules.

After all the data were collected, another experiment (Wright & Reid, 1973) was discovered which essentially duplicated the short-sentences and logical-tree conditions and, to a lesser extent, the retention conditions of the present experiment. Thus, the following experiment offers a check on the generality of findings for certain aspects of this previous investigation and also adds new factors in coding and variation in retention-related conditions.

METHOD

Subjects

Thirty male undergraduate students from Florida Technological University were randomly assigned, 10 to a group, to the three experimental conditions (short sentences, logical trees, and coding).

Materials

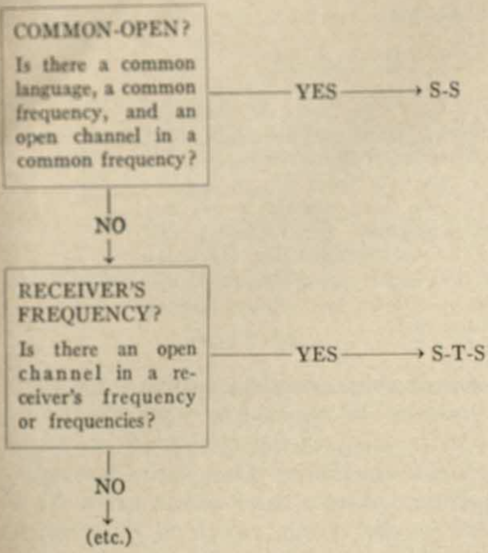
The apparatus consisted of a synthetic communications control console utilized by the subjects, and another control console with which the experimenter manipulated problems appearing on the subject's console. For each problem, the subject's task was to connect two hypothetical subscribers for communication purposes. This was accomplished by: (*a*) determining which of five modes of operation to use and (*b*) performing a routine (consisting of setting a series of buttons and dials) on the console appropriate to both the mode selected and particulars of the problem. Selection of modes was based on the experimentally manipulated instructions, on information contained in the problems, and on a wall chart. (See Bernstein and Gonzalez, 1971, for a detailed description of the experimental equipment and the subject's and experimenter's tasks.)

Tapes were played for each subject describing the characteristics of his console and its various modes of operation. The tapes differed among experimental groups in their reiteration of sets of printed instructions provided to the subjects. Printed instructions unique to the coding group are exemplified in the following:

---

[1] The author thanks Bernice Goetz, Ginger Sharkey, and John O'Hara for their assistance and comments.

Requests for reprints should be sent to Arthur S. Blaiwes, Naval Training Equipment Center, Code N-215, Department of the Navy, Orlando, Florida 32813.


COMMON-OPEN?
Is there a common language, a common frequency, and an open channel in a common frequency?
YES
S-S
NO
RECEIVER'S FREQUENCY?
Is there an open channel in a receiver's frequency or frequencies?
YES
S-T-S
NO
(etc.)

The short-sentences group received identical instructions in the following format: "*Use S-S if there is:* 1. A common language, 2. A common frequency, and 3. An open channel in a common frequency. *Use S-T-S if there is:* 1. No common language, and/or (etc.)." Logical-tree subjects received the same printed instructions given to coding subjects with the omission of the one- or two-word code preceding each question.

Two sets of thirty problems each were used equally for first and second test sessions. The order of sets of problems between test sessions and problems within sets was systematically varied among subjects. Among the six groups of five problems, problems were equated for difficulty and type both within each set and also between sets.

Procedure

For motivation purposes, the subjects were informed by the experimenter that they could earn from $2 to $4 per hour, depending on how well they performed, and were not informed of their performance level until their role in the experiment was completed. Subjects were treated individually in three distinct sections of the experiment.

During Section 1, each subject worked with only one of the presentation formats. Taped instructions were presented, and five minutes were allowed for study on the printed instructions. Then each subject was tested for five minutes at recalling his printed instructions and was corrected by the experimenter when in error. Next, following directions from the tape, the subject performed solutions on his console to one sample of each of the five problem types.

In Section 2, all subjects were presented with one set of 30 problems. Subjects were told they would be evaluated in terms of both speed and accuracy and that they could refer to their printed instructions after each problem was presented, but not after they started implementing their solution on the console. The experimenter informed the subjects when they were correct and told them about the nature of their errors when incorrect.

In Section 3, all subjects returned one week after the initial session for another session of 30 problems. These problems were presented in the same manner as those of the week before.

RESULTS AND DISCUSSION

No overall effects due to instructional format were found in separate analyses of variance on latency (times to initiate solutions to problems), accuracy (percentage of errors made in performing solutions to problems), or referrals (percentage of problems on which subjects referred to printed instructions). However, a significant interaction ($p < .05$) was obtained with all 60 problems on accuracy between instructional format and problem type (see Table 1). Individual Newman-Keuls tests showed that the coding group was significantly more accurate ($p < .01$) than both logical-tree and short-sentences groups on Deny problems. Further, accuracy scores of coding and short-sentences groups were significantly greater ($p < .01$ and $p < .05$, respectively) than those of the logical-tree group on S-S interrupt problems. Accuracy on the other three problem types did not differ significantly as a function of instructional format. These findings are important because Deny and S-S interrupt problems were more difficult to perform than the other three problem types (see Table 1).

Lack of accuracy differences on easier problems may have been due to a ceiling effect (i.e., the scores were incapable of measuring higher levels of ability). Alternately, maximal accuracy ability on the simple problems may not have required the more efficient formats. Absence of overall significance among accuracy scores (in view of the large accuracy differences across all problems) reflects the high variability within format groups (standard deviations on accuracy were 13%, 18%, and 21% for coding, short sentences, and logical trees, respectively).

Analyses of error types indicated that accuracy advantages among formats were mainly a function of differences in skill at implementing correct procedures once the proper mode was determined. Presumably, the more efficient formats served to unburden the performer from pressures of mode selection, allowing more time and thought for other aspects of the problem. A major error source contributing to the greater difficulty of Deny problems was a tendency to mistakenly use one of the "interrupt" modes instead of Deny.

Present data support the Wright and Reid finding that performance effects associated with presentation formats were manifested to the greatest extent when the problems were difficult. However, particulars from the two experiments differ under similar conditions; with difficult problems and written materials available to subjects, Wright's logical-tree group was more accurate than his short-sentences group. This discrepancy might be explained by the likelihood that present subjects operated from memory to a greater extent (see lower referral rates in Table 1) than subjects in Section I of the Wright study (where the logical-tree superiority was found). Where Wright's subjects were required to perform without the written instructions (Sections II and III), the superior logical-tree accuracy disappeared. Thus, when memory is required, it appears that logical trees are no better than short sentences, and in fact, short sentences may be better than logical trees. For logical trees to be superior to short sentences under memory-demanding conditions, coding may be required.

Another plausible explanation for the noted discrepancy in accuracy effects between the two experiments is that benefits typically derived from logical trees were lost in the present case due to confusion from the combining of conditions into compound sentences at single-decision points.

More rigorous training for retention in the present study may account for the finding that although accuracy and latency improved for some formats and deteriorated for others in the Wright study, these scores (along with referrals) consistently improved to the same extent with all formats in the present experiment. Also, if memory did fade, it could be refreshed under present conditions by referral to the instructions.

TABLE 1

MEAN SCORES FOR LATENCY, REFERRALS, AND ACCURACY AS A FUNCTION OF PRESENTATION FORMAT AND PROBLEM TYPE

Format	Problem type				
	Deny	S-S	S-S interrupt	S-T-S	S-T-S interrupt
Latency (in seconds)					
Short sentences	70^{a}	35^{b}	50^{c}	38^{d}	41^{c}
Logical tree	60	34	47	42	46
Coding	71	30	51	39	54
Referrals					
Short sentences	$.45^{a}$	$.09^{b}$	$.38^{c}$	$.22^{d}$	$.32^{c}$
Logical tree	.48	0	.23	.18	.19
Coding	.44	.04	.29	.22	.28
Accuracy (percentage of errors)					
Short sentences	42^{e}	9^{e}	14^{e}	12^{e}	10^{e}
Logical tree	43^{e}	9^{e}	20^{f}	15^{e}	11^{e}
Coding	24^{f}	9^{e}	11^{e}	11^{e}	8^{e}

Note. Results are for Sessions 1 and 2 combined.

[a,b,c,d] Cells within a score category having different superscripts are significantly different from each other.

[e,f] Means within a problem type having different superscripts are significantly different from each other.

Superior accuracy in the coding condition may have been due to several factors: (*a*) The codes may have served as aids to retrieval and/or reconstruction of a more coherent version of the rules. (*b*) The codes may have acquired the meaning of the rules resulting in faster and more accurate processing of rules. (*c*) Shorter times required to operate with the codes than with the rules, per se, may have permitted more practice in Section 1.

It is not clear from the present data whether this particular logical tree is a necessary condition to coding for the beneficial effects obtained from coding. That such shorthand languages probably do exert some influence on performance in a variety of contexts, however, is suggested by the fact that they develop naturally, over time, even where no systematic effort is made to do so. Further, coding is related to a notion concerning the general nature of retention. Theory suggests that "adult human memory may be derived from a reduced or telegraphic [coded] version of the material being retained . . . from which [the adult] . . . constructs the coherent recall [Cofer & Musgrave, 1963, p. 5]."

To the extent that the coding phenomenon is in fact general, it seems there should be an advantage to systematically assuring that the coding associated with a task is suitable and that it enters appropriately (in training and perhaps in a logical-tree format) into performance. Before such a recommendation can be made with confidence, however, further research is needed to unscramble possible interactions among coding, retention requirements, and other characteristics of the presentation format.

REFERENCES

Bernstein, B. R., & Gonzalez, B. K. *Learning, retention, and transfer*. (Tech. Rep. NAVTRADEVCEN 68-C-0215-1) Orlando, Fla.: Naval Training Device Center, 1971.

Cofer, C. N., & Musgrave, B. S. Introduction and summary. In C. N. Cofer & B. S. Musgrave (Eds.), *Verbal behavior and learning*. New York: McGraw-Hill, 1963.

Davies, I. K. Get immediate relief with an algorithm. *Psychology Today*, April [illegible], pp. 53-55; 68-69.

Jones, S. *Design of instruction*. London: Her Majesty's Stationery Office, 19[illegible].

Wright, P., & Reid, F. Written information: Some alternatives to prose for expressing the outcomes of complex contingencies. *Journal of Applied Psychology*, 1973, **57**, 160–166.

(Received January [illegible], 1974)

Journal of Applied Psychology
1974, Vol. 59, No. 6, 687–693

DESIGNING PREEMPLOYMENT TRAINING FOR THE HARD TO EMPLOY:

A CROSS-CULTURAL PSYCHOLOGICAL APPROACH[1]

HARRY C. TRIANDIS, JACK M. FELDMAN,[2] AND DAVID E. WELDON[3]
University of Illinois, Urbana-Champaign

WILLIAM M. HARVEY
St. Louis Narcotics Service Council

Most reports concerning the effects of training of the hard to employ suggest that the training is ineffective. One major factor responsible for the lack of success may be the neglect of cultural variables. A theoretical framework is presented which helps in the design of training programs that take cultural variables into account. Studies employing this framework are reviewed.

The last three years have produced a major focus of interest in the psychological literature toward understanding and solving the problems of the hard to employ (e.g., Feldman, 1973a, 1973b; Friedlander & Greenberg, 1971; Fromkin & Sherwood, 1974; Goodman, Paransky, & Salipante, 1973; Rosen & Turner, 1971; Triandis & Malpass, 1971). A conclusion of this research effort appears to be that training for the hard to employ is ineffective (Doeringer, 1969; Johnson, 1969; Morgan, Blonsky, & Rosen, 1970). Absenteeism, lateness, and turnover rates are unusually high among the hard to employ (Ferman, 1968; Purcell & Cavanaugh, 1972; Ross & Wheeler, 1971; Ruthledge & Gass, 1967) and do not appear to improve with training.

In the present article we argue that one of the major factors that is responsible for the lack of success of training programs is the neglect of cultural variables. Training which takes the cultural variables into account can be provided via the culture assimilator (Fiedler, Mitchell, & Triandis, 1971). A theoretical framework for the design of such training is presented.

PRELIMINARY CONSIDERATIONS

The behavior of "leaving the field" which is analogous to turnover, can be analyzed as a response that reduces noxious stimuli. Analyses of "culture shock" (Gullahorn & Gullahorn, 1963) suggest that a person who finds himself in a different culture sometimes experiences a severe disorientation and depression. His habitual responses do not result in reinforcements, but instead appear to be unrelated to outcomes. Without some preparation the experience can be so punishing that escape from the field is the only rewarding response.

A significant factor in reducing such disorientation is the development of institutions and personnel in the host environment who can recognize the visitor's needs and can help him adjust (Fromkin & Sherwood, 1974; Triandis, 1974b). Triandis and Malpass (1971) reviewed several factors which suggest that it is more important to train the supervisor of the hard to employ than it is to train the workers. In any case, this analysis shows the need for cultural training of *both* supervisors and workers.

The goal of intercultural training is to make the behavior of each participant in an interpersonal relationship more satisfying to the other member. Thus if the foreman and the worker belong to different cultures, good training should make the worker's behavior more reinforcing to the foreman and the foreman's behavior more reinforcing to the worker. If this is accomplished the worker will be less likely to leave the field.

---

[1] This investigation was supported by Research Grant 15-P-55175/5 from the Social and Rehabilitation Service of the Department of Health, Education, and Welfare (H. C. Triandis, principal investigator).

Requests for reprints should be sent to Dr. Harry C. Triandis, Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, Illinois 61820.

[2] Now at the University of Florida, Gainesville, Florida.

[3] Now at Washington University, St. Louis, Missouri.

ANALYSIS OF CULTURAL VARIABLES

The classic anthropological approach is participant observation. One lives in the community he is to study and interviews various "informants" about the way things happen. An ethnography is a report of the way people use language, their customs, privileges and duties in relation to various types of people in their community, forms of social and political organization, economic activity, belief systems, religious rituals, values, and so on. In the case of the black ghetto good material reflecting such topics can be found in Clark (1965), Gladwin (1967), Hannerz (1969), Keil (1966), Liebow (1967) and Suttles (1968). The most perceptive presentation and summary is probably Valentine's (1968).

The black ghetto is different from other black and white environments and results in perceptions of the social environment that differ substantially from those typically found among the middle class. As an example consider some differences in language:

Black Ghetto English[4]	Standard English
ace	= good friend
bad	stylish, very good, or boss
down	agreeable, best ever
F.B.I.	fat, black, and ignorant
fuzz	police
gig	job

In addition, there is evidence of cultural differences in attitudes toward nonstandard family patterns (Rodman, Nichols, & Voydanoff, 1969), greater latitude for what is considered acceptable behavior (Rodman, 1963) and weaker perceived connections between behavior and outcomes (Triandis, Feldman, & Harvey, 1970, 1971a, 1971b, 1971c; Triandis, Weldon, & Feldman, 1972) in the ghetto than in the middle class, black or white. The ghetto is characterized by particular perceptions of social relations and ecosystem distrust (Triandis, Feldman, Weldon, & Harvey, in press). Ecosystem distrust is a label that subsumes such phenomena as lower interpersonal trust, rejection of institutions in one's environment, suspicion of authority figures, and the perception of little or no connection among events occurring in the environment.

It seems obvious that if such differences are not taken into account when designing training programs, it is very likely that the training will prove unsuccessful. To improve interpersonal relations between blacks and whites in job settings, we need to train whites (*a*) to recognize the substantial differences which exist among different kinds of blacks (e.g., ghetto versus middle class), (*b*) to understand the black point of view. Similarly, we must train the black ghetto workers to understand the middle-class point of view, black or white.

Differences between blacks and whites, and the poor and the rich, have been discussed for a long time; the literature is voluminous. Race differences have been discussed in Baughman (1971), Ehrlich (1973), Fanon (1967), Frazier (1932, 1966), Glazer and Moynihan (1963), Goldschmid (1970), Hauser (1971), Jones (1972), Lester (1968), Miller and Dreger (1973), Myrdal (1944), Pettigrew (1964), Rose (1948), and Wilcox (1971), among many others. Social class differences have been discussed extensively by Lewis (1959, 1960, 1961, 1966) and Miller (1958). Difficulties associated with the employment of blacks have been discussed by Hearns (1968), Kirkpatrick (1973), Pinto and Buchmeier (1973), and Purcell and Cavanaugh (1972), among others. Obviously, there are practical difficulties in imparting this much information during training. We need a framework which selects those aspects of the information that are most important for the improvement of interpersonal relations between a black ghetto worker and his supervisors, black or white. This framework is presented below.

---

[4] Personal communication from Dr. Daniel Landis.

THEORETICAL FRAMEWORK FOR THE DESIGN OF CROSS-CULTURAL TRAINING

Triandis (1974a) has analyzed the problem of interpersonal behavior across cultures as one in which people make inappropriate attributions (Kelley, 1967). A relatively well established phenomenon discussed by attribution theorists (Jones & Nisbett, 1971) is that the *actor* sees his own behavior as controlled mostly by situational factors, while the *observer* sees the actor's behavior controlled mostly by personality traits or internal predispositions. Kelley puts it slightly differently, pointing out that the observer does not give sufficient weight to situational variables. Triandis (1974a) argues that one of the ways to improve interpersonal relations is to make the observer more similar to the actor in his attributions about the causes of the actor's behavior. In short, to increase *isomorphic attributions*. The actor is usually a more accurate analyst of his own behavior (except in the case of abnormal behavior). Thus, by using the criterion of isomorphic attributions, we can improve the effectiveness of the observer's analysis of the actor's behavior. Furthermore, when the observer makes isomorphic attributions, he increases his general similarity to the actor, which should eventually lead to greater interpersonal attraction (Byrne, 1969).

For example, if a supervisor and a subordinate agree that a job is to be finished by a certain date, but the supervisor perceives the antecedents of this event as "employee participation," while the subordinate perceives it as "the boss's orders," when the employee later discovers that he is unable to complete the job on time, the supervisor may conclude that the employee is "unable to set realistic time standards" while the employee may conclude that "the boss gave the *wrong* orders," leading to a cycle of misunderstandings. Isomorphic attributions would involve either both agreeing that it is an order or both understanding the principles of employee participation.

Triandis's (1974a) analysis of interpersonal conflicts across cultures suggests that nonisomorphic attributions are a major, though a subtle, problem. It is particularly serious in cross-cultural encounters because people have different "subjective cultures."

Culture is here defined as the man-made part of the human environment (Herskovits, 1955). Subjective culture is the subjective reaction of members of a cultural group to the man-made part of the environment. A cultural group is a set of individuals who speak a mutually understandable dialect (Naroll, 1968). Subjective culture can be studied through the development of culture-specific (emic) measures which have culture-general (etic) properties, as described in Triandis, Vassiliou, Vassiliou, Tanaka, and Shanmugam, (1972). Such an approach permits each cultural group to express itself in a manner that is most appropriate for it, but also permits comparisons of the responses of different cultural groups.

Social scientists have used about 80 concepts that can be defined as intervening variables, mediating between consistencies in social situations on the one hand and consistencies in responses on the other (Campbell, 1963). Among the most frequently used of these concepts, listed alphabetically, are attitudes, beliefs, concepts, norms, roles, and values. Triandis, Vassiliou, Vassiliou, Tanaka, and Shanmugam, (1972) argue that some of these concepts are required for the prediction of behavior, but a much smaller set of concepts than the set currently in use is essential. They present a theoretical framework which uses (*a*) behavioral intentions, measured with the behavioral differential (Triandis, 1964); (*b*) norms; (*c*) roles; (*d*) general behavioral intentions, measured with norm and role differentials (Triandis, Vassiliou, & Nassiakou, 1968) and generalized behavioral differentials; (*e*) the affect toward the behavior, measured by semantic differentials (Osgood, Suci, & Tannenbaum, 1957) in which the behavior is used as a concept and it is rated on several scales of the evaluative factor of the semantic differential; (*f*) the perceived consequences of the behavior, measured on probability scales with a semantic differential format; and (*g*) the value of the consequences, measured by semantic differentials in which the consequence is the concept rated on evaluative scales.

Specifically, but in summary, Triandis argues that the probability of a particular behavior (P_a) is a function of previously established habits (H), the intention to behave (I), and facilitating conditions (F).

$$P_a = (\alpha H + \beta I)F.$$

Habits reflect the subject's previous reinforcement history; facilitating conditions include motivation, ability, and the difficulty of carrying out the act. Intentions are a function of social factors (S), the affect toward the behavior (A), and the cognitions (C) associated with the behavior.

$$I = \gamma S + \delta A + \epsilon C.$$

The social factors reflect existing norms (N), roles (R), the self-concept (I am the kind of person who behaves this way), and interpersonal agreements or contracts; the affect toward the behavior reflects classical conditioning involving the thought of the behavior, resulting from pleasant or unpleasant experiences associated with the particular behavior; the cognition is a function of the perceived consequences and the value of these consequences.

$$C = \sum_{c=1}^{n} P_c V_c.$$

The coefficients α, β, γ, δ, and ϵ reflect interactions among characteristics of the person, situation, and behavior. For example, high *F*-scale subjects, in cultures that clearly specify that a person must pray in a particular way at a particular point in time each day, typically do so. This indicates a very high γ coefficient for praying when the cues suggest it is the proper time. High coefficients for one variable imply low coefficients for the others, because

$$\alpha + \beta = 1 \text{ and } \gamma + \delta + \epsilon = 1.$$

When a person from one culture observes the behavior of a person from another culture, he is likely to make wrong attributions because of differences in (*a*) the weights of a particular determinant of behavior and (*b*) the values of H, I, F, S, A, or C.

The analysis of subjective culture (Triandis, Vassiliou, Vassiliou, Tanaka, & Shanmugam, 1972) requires the acquisition of data relevant to each of these points. Thus, for example, one studies the conceptions of proper behavior in role situations (Triandis et al., 1968). Triandis et al. (1970, 1971a, 1971b, 1971c) covered most of these points for ghetto blacks and Triandis, Weldon, and Feldman (1972) cross-validated the findings with another sample of the hard to employ. From such data one can discover where the attributions made by members of one culture are likely to be most inaccurate, when they perceive the behavior of members of another culture.

One kind of training should be directed toward changing these attributions. Specifically, let us assume that members of Culture A see the behavior of a member of Culture B and attribute it to his "motivation," while members of Culture B looking at the very same behavior attribute it to his perceived role. It seems plausible that members of Culture B, who after all know more about the causes of the behavior of members of their own culture, would be more accurate in their analysis of the causes of this person's behavior. Thus, if we train people from Culture A to see the behavior as due to role perceptions, we might improve their ability to work with the particular person. This is the basic principle behind the most recent "culture assimilators."

Earlier assimilators seemed to have been somewhat effective (Fiedler et al., 1971; Mitchell, Dossett, Fiedler, & Triandis, 1972), but they were not based on a theoretical analysis. The new assimilators should prove more effective.

The culture assimilator is a programmed learning sequence in which incidents of intercultural conflict are described and the trainee is asked to judge the relevance of various factors as causes of the incident. Thus, a one or two paragraph story is presented, which is followed by a multiple-choice format with four answers. When training members of Culture A, one of the answers is an attribution that members of Culture B make in explaining the behavior of members of Culture B. The other three are attributions that are typically made by members of Culture A to explain the behavior in the particular incident. After the trainee selects one of the four answers, he is given feedback. If he selects the right answer he is told why it is the right answer, usually by reference to norms, values, etc. of the other group. If he gives the wrong answer, he is told to read the incident again and selects another answer.

In training white or black middle-class managers we might use an incident reflecting ecosystem distrust. In the hypothetical incident a foreman tells his black ghetto worker that if the worker behaved in a particular way he would receive a 10-cent-an-hour raise. Some time passes during which the worker appears unresponsive.

Explanations of the unresponsiveness of the worker might include: "He is lazy," "He is dumb," "He is not interested in more money," and "He does not trust his foreman." Given the information that the worker is from the black ghetto and what we have found about ecosystem distrust, the last attribution is more probable than the others. Training the supervisor to make such attributions is likely to suggest to him other ways of handling the situation. For example, he might make a "promise" that requires the worker to do something that is very easy for him to do and promptly fulfill his promise. Repeated exposure of the worker to the connection "supervisor is trustworthy" may make other promises, which are less easy to fulfill, more believable.

The relationship between the black worker and his job environment may involve literally thousands of "interactions." Each interaction can be misunderstood; distrust can give it a different meaning. The negative meaning of such interactions produces a total impact of a "very negative environment," hence the "leaving the field" behavior. The hypothesis suggested by this analysis is that training the supervisor (and the fellow workers) may result in enough changes in the valence of these interactions to reduce the noxiousness of the environment for the black ghetto worker.

DISCUSSION

The approach advocated here should not be considered a panacea. In fact, it is too early to know how well it is going to work. So far we have established only that trained individuals change the kinds of attributions they make about the behavior of others from other cultures, and this allows them to be more "tolerant" in judging such behavior (Weldon, Carlston, Rissman, Slobodin, & Triandis, 1974). But we are still at the level of affective responses, not at the level of changes in behavior. One could argue that changing the way one feels about a culture is a sine qua non, if other desirable changes are to occur. But this remains to be established.

It is unlikely that the assimilator will work in situations in which the person is not reinforced by his membership groups for "appropriate" behavior toward members of other cultures. Thus, if the supervisor of a foreman that is trained with the assimilator rewards the foreman who retains more hard to employ workers, and punishes, or at least shows his displeasure toward, the foreman who has poor retention rates, this "environment" is likely to make the assimilator useful. An environment in which management is not providing differential reinforcements will not allow the assimilator to do his part in cross-cultural training.

When comparing the assimilator with other methods of training it should be remembered that it is a highly concentrated form of "cognitive training." Basically, all it does is impart information. One could accomplish the same aim with a book. However, many trainees find programmed learning materials of a "problem-solving" kind, such as the assimilator, more enjoyable than a book.

It is difficult to compare the assimilator with T groups and encounter groups because there are so many different kinds of such groups (Back, 1974). However the one point that is clear is that in such groups a person might make cross-cultural errors, in judgment or behavior, which may prove embarrassing. The assimilator is "safe" from embarrassment. It seems reasonable to assume that assimilator training should come before other kinds of training to protect the trainee from embarrassment. Comparisons are also difficult because assimilator training takes only 6–10 hours, while most T groups take much more time.

Role playing the other group's roles may be helpful, but role playing assumes the trainee knows how to take the role of the other. If he has no information, his role taking may be silly. Again, the culture assimilator may be best *before* such playing.

In addition to cognitive training there are forms of training which might be labeled "affective" (e.g., having pleasant experiences in the presence of members of the other culture) and "behavioral" (e.g., learning to emit behaviors which members of the other culture find rewarding). There is no contradiction among these types of training, and some combination of them should prove superior to a single method. It is likely that a behavior modification approach added to assimilator training will prove particularly effective. Here the psychologist identifies behaviors that each cultural group finds rewarding and arranges "contracts" for each to reward the other and for the mutual reinforcement of a supervisor–subordinate pair.

The cultural information obtained to develop the assimilators can be used also in the development of such a behavior modification program. Thus, this information is fundamental to any attempts at integration of the anthropological and psychological approaches to training.

REFERENCES

Back, K. W. Intervention techniques: Small groups. *Annual Review of Psychology*, 1974, **25**, 367–387.

Baughman, E. E. *Black Americans: A psychological analysis*. New York: Academic Press, 1971.

Byrne, D. Attitudes and attraction. In L. Berkowitz (Ed.), *Advances in experimental psychology*. Vol. 4. New York: Academic Press, 1969.

Campbell, D. T. Social attitudes and other acquired behavioral dispositions. In S. Koch (Ed.), *Psychology: A study of a science*. New York: McGraw-Hill, 1963.

Clark, K. B. *Dark ghetto: Dilemmas of social power*. New York: Harper, 1965.

Doeringer, P. B. *Programs to employ the disadvantaged*. Englewood Cliffs, N.J.: Prentice-Hall, 1969.

Ehrlich, H. J. *The social psychology of prejudice*. New York: Wiley, 1973.

Fanon, F. *Black skin, white masks*. New York: Grove Press, 1967.

Feldman, J. M. Race, employment, and the evaluation of work. *Journal of Applied Psychology*, 1973, **58**, 10–15. (a)

Feldman, J. M. Race, economic class, and perceived outcomes of work and unemployment. *Journal of Applied Psychology*, 1973, **58**, 16–22. (b)

Ferman, L. A. *The Negro and equal employment opportunities*. New York: Praeger, 1968.

Fiedler, F. E., Mitchell, T., & Triandis, H. C. The culture assimilator: An approach to cross-cultural training. *Journal of Applied Psychology*, 1971, **55**, 95–102.

Frazier, E. F. *The Negro family in Chicago*. Chicago: University of Chicago Press, 1932.

Frazier, E. F. *The Negro family in the United States*. (Rev. and abridged ed.) Chicago: University of Chicago Press, 1966.

Friedlander, F., & Greenberg, S. Effect of job attitudes, training, and organization climate on performance of the hard-core unemployed. *Journal of Applied Psychology*, 1971, **55**, 287–295.

Fromkin, H. L., & Sherwood, J. *Integrating the organization*. Glencoe, Ill.: Free Press, 1974.

Gladwin, T. *Poverty, U.S.A.* Boston: Little, Brown, 1967.

Glazer, N., & Moynihan, D. P. *Beyond the melting pot: The Negroes, Puerto Ricans, Jews, Italians, and Irish of New York City*. Cambridge: M.I.T. Press and Harvard University Press, 1963.

Goldschmid, M. L. *Black Americans and white racism: Theory and research*. New York: Holt, Rinehart & Winston, 1970.

Goodman, P. S., Paransky, H., & Salipante, P. Hiring, training, and retaining the hard-core unemployed: A selected review. *Journal of Applied Psychology*, 1973, **58**, 23–33.

Gullahorn, J. T., & Gullahorn, J. E. An extension of the U-curve hypothesis. *Journal of Social Issues*, 1963, **19**, **3**, 33–47.

Hannerz, U. *Soulside: Inquiries into ghetto culture and community*. New York: Columbia University Press, 1969.

Hauser, S. T. *Black and white identity formation*. New York: Wiley, 1971.

Hearns, J. P. New approaches to meet post-hiring difficulties of disadvantaged workers. In G. C. Summers (Ed.), *Industrial Relations Research Association: Proceedings of the 21st annual winter meetings*, 1968, 207–216.

Herskovits, M. J. *Cultural anthropology*. New York: Knopf, 1955.

Johnson, L. A. *Employing the hardcore unemployed. AMA research study 98*. New York: American Management Association, 1969.

Jones, E. E., & Nisbett, R. E. *The actor and the observer: Divergent perceptions of the causes of behavior*. New York: General Learning Press, 1971.

Jones, R. L. *Black psychology*. New York: Harper & Row, 1972.

Keil, C. *Urban blues*. Chicago: University of Chicago Press, 1966.

Kelley, H. H. Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska Symposium on Motivation: 1967*. Lincoln: University of Nebraska Press, 1967.

Kirkpatrick, J. J. Occupational aspirations, opportunities and barriers. In K. S. Miller & R. M. Dreger (Eds.), *Comparative studies of blacks and whites in the United States*. New York: Seminar Press, 1973.

Lester, J. *Look out, whitey! Black power's gon' get your mama!* New York: Grove Press, 1968.

Lewis, O. *Five families: Mexican case studies in the culture of poverty.* New York: Basic Books, 1959.

Lewis, O. *Tepoztlan: Village in Mexico.* New York: Holt, 1960.

Lewis, O. *The children of Sanchez.* New York: Random House, 1961.

Lewis, O. The culture of poverty. *Scientific American,* 1966, 215 (4), 19–25.

Liebow, E. *Tally's corner. A study of Negro streetcorner men.* Boston: Little, Brown, 1967.

Miller, K. S., & Dreger, R. M. *Comparative studies of blacks and whites in the United States.* New York: Seminar Press, 1973.

Miller, W. B. Lower class culture as a generating milieu of gang delinquency. *Journal of Social Issues,* 1958, 14, 5–19.

Mitchell, T., Dossett, D. L., Fiedler, F. E., & Triandis, H. C. Culture training: Validation evidence for the culture assimilator. *International Journal of Psychology,* 1972, 7, 97–104.

Morgan, B. S., Blonsky, M. R., & Rosen, H. Employee attitudes toward a hardcore hiring program. *Journal of Applied Psychology,* 1970, **54**, 473–478.

Myrdal, G. *An American dilemma.* New York: Harper, 1944.

Naroll, R. Some thoughts on comparative method in anthropology. In H. M. Blalock & A. B. Blalock (Eds.), *Methodology in social research.* New York: McGraw-Hill, 1968.

Osgood, C. E., Suci, G., & Tannenbaum, P. *The measurement of meaning.* Urbana, Ill.: University of Illinois Press, 1957.

Pettigrew, T. F. *A profile of the Negro American.* Princeton: Van Nostrand, 1964.

Pinto, P. R., & Buchmeier, J. O. Problems and issues in the employment of minority disadvantaged and female groups: An annotated bibliography. (Bulletin 59). Minneapolis: Industrial Relations Center, University of Minnesota, 1973.

Purcell, T. V., & Cavanaugh, G. F. *Blacks in the industrial world: Issues for the manager.* New York: Free Press, 1972.

Rodman, H. The lower class value stretch. *Social Forces,* 1963, 42, 205–215.

Rodman, H., Nichols, F. R., & Voydanoff, P. Lower-class attitudes toward "deviant" family patterns: A cross-cultural study. *Journal of Marriage and the Family,* 1969, **31**, 315–321.

Rose, A. *The Negro in America.* Boston: Beacon Press, 1948.

Rosen, H., & Turner, J. Effectiveness of two orientation approaches in hard-core unemployed turnover and absenteeism. *Journal of Applied Psychology,* 1971, **55**, 296–301.

Ross, J. C., & Wheeler, R. H. *Black belonging.* Westport, Conn.: Greenwood Press, 1971.

Ruthledge, A. L., & Gass, G. Z. *Nineteen Negro men.* San Francisco: Jossey-Bass, 1967.

Suttles, G. D. *The social order of the slum: Ethnicity and territory in the inner city.* Chicago: University of Chicago Press, 1968.

Triandis, H. C. Exploratory factor analyses of the behavioral component of social attitudes. *Journal of Abnormal and Social Psychology,* 1964, **68**, 420–430.

Triandis, H. C. Culture training, cognitive complexity, and interpersonal attitudes. In R. W. Brislin, S. Bochner, & W. J. Lonner (Eds.), *Cross-cultural perspectives and learning.* Vol. 1. Beverly Hills, Calif.: Sage, 1974. (a)

Triandis, H. C. Person perception: A review of the literature and implications. In H. L. Fromkin and J. Sherwood (Eds.), *Integrating the organization.* Glencoe, Ill.: Free Press, 1974. (b)

Triandis, H. C., Feldman, J., & Harvey, W. Person perception among black and white adolescents and the hardcore unemployed. (Tech. Report No. 5, SRS No. RD 2841-G) Champaign, Ill.: Department of Psychology, University of Illinois, 1970.

Triandis, H. C., Feldman, J., & Harvey, W. Job perceptions among black and white adolescents and the hardcore unemployed. (Tech. Rep. No. 7, SRS No. RD 2841-G) Champaign, Ill.: Department of Psychology, University of Illinois, 1971. (a)

Triandis, H. C., Feldman, J., & Harvey, W. The perceptions of implicative relationships among black and white adolescents and the hardcore unemployed. (Tech. Rep. No. 8, SRS No. RS 2841-G) Champaign, Ill.: Department of Psychology, University of Illinois, 1971. (b)

Triandis, H. C., Feldman, J., & Harvey, W. Role perceptions among black and white adolescents and the hardcore unemployed. (Tech. Rep. No. 6, SRS No. RD 2841-G) Champaign, Ill.: Department of Psychology, University of Illinois, 1971. (c)

Triandis, H. C., Feldman, J., Weldon, D., & Harvey, W. Eco-system distrust and the hard to employ. *Journal of Applied Psychology,* in press.

Triandis, H. C., & Malpass, R. S. Studies of black and white interaction in job settings. *Journal of Applied Social Psychology,* 1971, **1**, 101–117.

Triandis, H. C., Vassiliou, V., & Nassiakou, M. Three cross-cultural studies of subjective culture. *Journal of Personality and Social Psychology,* 1968, **8** (4, Pt. 2), 1–42.

Triandis, H. C., Vassiliou, V., Vassiliou, G., Tanaka, Y., & Shanmugam, A. *The analysis of subjective culture.* New York: Wiley, 1972.

Triandis, H. C., Weldon, D. E., & Feldman, J. Black and white hardcore and middle class subjective cultures: A cross validation. (Tech. Rep. No. 14, SRS 15-P-55175/5) Champaign, Ill.: Department of Psychology, University of Illinois, 1972.

Valentine, C. A. *Culture and poverty: Critique and counter-proposals.* Chicago: University of Chicago Press, 1968.

Weldon, D., Carlston, D., Rissman, A., Slobodin, L., & Triandis, H. C. A laboratory test of the effects of culture assimilator training. (Tech. Rep. No. 20, SRS 15 P-55175/5) Champaign, Ill.: Department of Psychology, University of Illinois, 1974.

Wilcox, R. C. *The psychological consequences of being a black American.* New York: Wiley, 1971.

(Received December 26, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 6, 694-698

BEHAVIOR MODIFICATION AND ABSENTEEISM: INTERVENTION IN ONE INDUSTRIAL SETTING[1]

ED PEDALINO[2]
Pedalino and Associates, Inc., Ann Arbor, Michigan

VICTOR U. GAMBOA
School of Business Administration, University of Michigan

Behavior modification was used in an attempt to decrease absenteeism in a sample of 215 hourly employees at a manufacturing/distribution facility. Employees in four adjoining plants served as comparison groups. An ABA (baseline, intervention, return to baseline) intervention using a lottery incentive system constituted the experimental design. Absenteeism decreased significantly following the experimental group intervention but did not decrease in any of the four comparison groups. Further, stretching the schedule of reinforcement did not increase the rate of absenteeism. Findings are discussed in light of Lawler and Hackman's 1969 study which indicated participation, not the incentive system, decreased absenteeism.

There is a growing interest in the application of behavior modification in industrial organizations. One area of potential application has been the absenteeism of the marginal worker (Porter, 1973). Interest in the use of behavior modification is on two levels—theoretical and applied. On a theoretical level, two landmark articles have been written, Porter (1973, pp. 113–133) and Nord (1972). Also Gamboa (1974) addressed the issue of compatibility between behavior modification and current organizational theories of motivation, and came to the conclusion that the two can be synthesized along content and process dimensions. Jablonsky and DeVries (1972) critically evaluated Nord's article and developed an "interactive" model of behavior which accounts for multiple reinforcement on employee behavior.

On the applied level, the laboratory experiments of Yukl, Wexley, and Seymore (1972), Cherrington, Reitz, and Scott (1971), and Adam (1972) attempted to measure the efficacy of behavior modification in addressing organizational issues (i.e., pay incentives, satisfaction–productivity quagmire, increasing performance quality and quantity). Popular literature citing successful application of behavior modification in industry has largely revolved around the Emery Air Freight experience ("New Tool," 1971; "Where Skinner's Theories Work," 1972). Nord (1972) described a case study of a hardware company which used a lottery system to decrease absenteeism and tardiness. Brethower (1973), in a field study, focused on the use of feedback as a behavioral technique to intervene in two industrial settings. With the exception of Brethower, there is a dearth of field experiments evaluating the applicability of behavior modification in industrial organizations.

The article written by Porter (1973) provided the theoretical and applied framework on which to base this study. Porter's focus is on the marginal worker, defined as an individual who has failed to demonstrate consistent work attendance and/or has failed to meet organizationally defined standards of adequate performance (Porter, 1973). According to Porter, acquisition of behavior via shaping, and maintenance of acquired behavior via schedules of reinforcement are the key tools for improving attendance and performance on the part of the marginal worker.

The basic hypothesis of this study is that behavior modification can be utilized in an industrial organization to reduce absenteeism. Gamboa (1974) reviewed the literature on absenteeism and concluded that the bulk of studies reported were either surveys or anecdotal reports of what companies did to decrease absenteeism. Three studies, Mann and

---

[1] This paper was originally presented at the Academy of Management Convention at Seattle, Washington, August 1974.

[2] Requests for reprints should be sent to Ed Pedalino, Pedalino and Associates, Inc., 4965 Pitchpine Court, 1B, Ann Arbor, Michigan 48197.

Baumgartel (1952), Mann and Sperling (1956), and Lawler and Hackman (1969) were exceptions. All were interventions in industrial settings and attempted to evaluate the effects of their intervention on absenteeism.

The study by Lawler and Hackman (1969) is related to the present study in two important aspects: (*a*) the use of a pay incentive plan to decrease absenteeism and (*b*) the use of a field experiment to evaluate the effect of experimental intervention. Lawler and Hackman (1969), however, focused on participation of employees as the explanatory variable leading to decreased absenteeism. The present study focused on the use of behavior modification and attempts to test its efficacy in decreasing absenteeism.

B. F. Skinner, in an interview, has only recently stated his ideas on how behavior modification can be applied in industrial organizations (see Conversation with Skinner, 1973). For example, Skinner suggests that a lottery (such as a door prize) may well solve the problem of absenteeism. Formally stated, therefore, the basic hypothesis of the study is: intermittent reinforcement using financial rewards will lead to a significant decrease in absenteeism.

METHOD

Research Strategy

The basic hypothesis of the study was tested in a field experiment. The experimental approach chosen followed a two-pronged design: (*a*) the ABA design whereby baseline measures are taken (A), intervention takes place (B), and the intervention strategy is removed (A) with measures systematically monitored in all three phases, and (*b*) the "controlled" design whereby the experimental group is compared with four other groups in the same research site; the "control" feature uses comparison groups to assess the impact of trend as an alternative hypothesis for any change in absenteeism rate.

Research Site and Subjects

The research was conducted at a manufacturing/distribution center that handles numerous products, with each product housed in a separate plant. Each plant operation was, for all practical purposes, independent of the others. Experimental intervention took place in one plant with the other plants serving as comparison groups.

The plants were made up of a four-level reporting hierarchy with a plant manager at the top, then superintendents, followed by foremen, with unionized employees at the lowest level. Each foreman was responsible for a group of employees ranging in size from 14 to 26.

The average employee had an eleventh grade education. Only full-time employees (95% of total) were included in the study. All were blue-collar unionized employees. The average amount of time an employee worked in the experimental plant was approximately one year and six months. All were hired through the division personnel office, where applicants were required to be at least 18 years or older, with no felony or narcotics record, and in acceptable physical health. The average wage received was $4.86 per hour paid on a weekly basis. Four of the five groups followed the same sick day policy (allowed six days a year of paid sick leave); the only exception was Group 4.

The experimental group and Group 1 were assembly line operations, while Groups 2 and 3 were order picking and unloading operations, and Group 4 involved unloading and meat processing. Groups 1 to 4 had more than one shift. Size of the groups ranged from 215 (experimental) and 370 (Group 2) to 65 (Group 3), 43 (Group 4) and 30 (Group 1).

Prior to the intervention, the plant chosen to be the experimental group had a significantly lower absenteeism rate compared to the other plants. This plant was chosen as the experimental group since the plant manager initiated the request to reduce absenteeism, thus reducing usual entry and resistance difficulties. However, matching or random assignment of subjects did not take place. Thus, rather than being "control" groups in the strict sense, the other four groups would be more accurately described as "comparison" groups.

Procedure

The behavior modification procedure utilized to reduce absenteeism consisted of four phases:

1. Baseline 1—where the problem was identified in measurable terms.
2. Phase 1—where the program was operational *each* week for six weeks.
3. Phase 2—where the program was operational *every other* week for ten weeks.
4. Baseline 2—where the program was removed.

Baseline 1. Absenteeism records for the five groups in the company were obtained in order to determine baseline attendance measures. Other than for vacation, bereavement, jury duty, and paid holidays, any employee not present for work was considered absent. An absenteeism rate was computed weekly by determining the number of man-days lost and comparing this to total number of man-days possible. Such baseline data was obtained for 32 weeks for all groups prior to installing the intervention program.

Phases 1 and 2. A lottery incentive system was developed in line with two basic principles of behavior modification—shaping and stretching the schedules of reinforcement. From the lottery notion, the idea of using a poker game intervention strategy occurred. The poker game idea was checked with management and their reaction was positive. Thus, a poker game incentive plan was chosen as the intervention strategy. The plan was to work as follows:

> Each day an employee *comes* to work and is *on time,* he is allowed to choose a card from a deck of playing cards. At the end of the five-day week, he will have five cards or a normal poker hand. The highest hand wins $20. There will be eight winners, one for approximately each department.

The implementation of such a strategy constituted Phases 1 and 2 of the program. Basically the program provided an added incentive for entering the work environment each day.

Since the number of men in each department ranged from 14 to 26, the chances of winning were greater in certain departments than in others. However, since each department was a naturally existing functional unit, it was decided to preserve the current divisions for the program.

Each time an employee came to work on time, he received a card from a poker deck from his foreman. The foreman recorded the card chosen by the employee on a score sheet. The foreman also recorded the card on a poker game chart. The chart was a larger version of the score sheet and was posted where men of the same department could see how everyone was doing.

It is one thing to introduce a program to reduce absenteeism, and it is another thing to maintain such reduced absenteeism over a period of time. A fading procedure, therefore, was designed to stretch the interval of reinforcement. In Phase 1, the first 6 weeks, the program was run each week (fixed interval—1 week). In Phase 2, the last 10 weeks, the program was run every other week—fixed interval—2 weeks).

Baseline 2. After 16 weeks the program was discontinued. Absenteeism measures over 22 weeks continued to be monitored for the experimental group to determine long-run effectiveness of the program. The procedure employed provides two ways of evaluating the effectiveness of the experimental intervention: (*a*) the ABA design uses the experimental group as its own control in determining the differential impact of introducing and removing the intervention, and (*b*) the comparison groups enable us to "control" for *trend* as an alternative hypothesis to explain the effect of experimental intervention.

RESULTS

Results of the experiment are shown in Figure 1. The baseline measure of 32 weeks showed the overall weekly absenteeism rate of the experimental group to be 3.01%. The goal aimed for was an absenteeism rate of 2.31%. This was determined by computing what the absenteeism rate would be if all employees took just their six days sick leave each year. It should be noted that absenteeism rates for some weeks had been as low, if not lower, than the 2.31% goal line. In effect, with existing personnel, technology, and structure, the goal of 2.31% had already been achieved and was thus realistic.

Figure 1 shows that the overall absenteeism rate actually achieved as a result of the intervention was 2.46%. Compared to the baseline of 3.01% this represented an 18.27% decrease in absenteeism. This baseline-intervention decrease in absenteeism was tested for statistical significance by a one-tailed t test ($t = 1.98$, $p < .05$). A one-tailed test was chosen since a directional decrease in absenteeism from baseline to intervention period was hypothesized for the experimental group. A 22-week follow-up after the incentive system intervention was phased out showed the absenteeism rate to have climbed to 3.02%. This represented an increase of 30.08% over the intervention rate. This intervention–follow-up increase was tested for statistical significance by a two-tailed t test ($t = 3.34$, $p < .01$). A two-tailed test was chosen since there was no a priori basis for hypothesizing a change in absenteeism in a specific direction for this intervention–follow-up period.

An unplanned manipulation occurred in the eighth week of the experimental period. In Phase 1, the incentive program was structured to reward good attendance (five work days present on time) on a weekly basis; in Phase 2, the incentive program was to be operational on a biweekly basis. Thus, while the bonus for the poker game was awarded on the week of September 24, no poker game was to take place on the week following. The next poker game was to take place on October 15, and so on.

Unanticipated circumstances resulted in no incentive program being put into operation on the weeks of October 1, 8, and 15, three weeks in a row. Absenteeism, as can be shown in Figure 1, immediately increased from 2.3% to 3.9%. The interval for reinforcement had been stretched beyond the point which could maintain good attendance. When the incentive was again reinstituted, absenteeism immediately dropped to 2.4% and stayed at about that level for the next six weeks. In

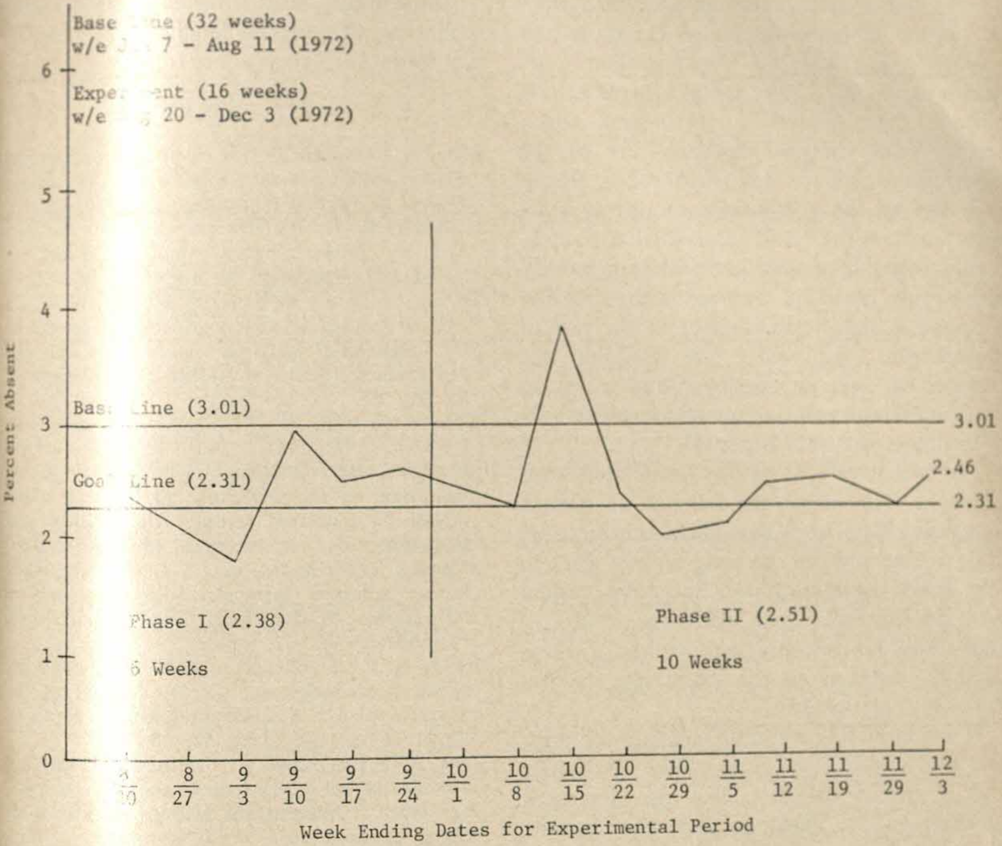


FIGURE 1. Experimental group attendance program results.

effect, an accidental manipulation within the ABA design provided further evidence for the efficacy of the intervention.

To test the effectiveness of "stretching the schedule of reinforcement," the average absenteeism rate of the employees in the experimental group during Phase 1 (weekly schedule of reinforcement) and Phase 2 (biweekly schedule of reinforcement) were compared. During Phase 1, the average employee in the experimental group was absent 2.38% of total scheduled work hours. During Phase 2 the employee was absent 2.51% of total scheduled work hours. A two-tailed *t* test showed no significant difference between absenteeism for Phase 1 and Phase 2.

The baseline–intervention data of the experimental group vis-à-vis the comparison groups were then examined. Over the same 16-week intervention period, matching of baseline and intervention measures showed an increase in absenteeism in four of the comparison groups. Such an intervention–baseline increase was tested for statistical significance by a one-tailed *t* test, and only one of the comparison groups (Group 1) yielded a *t* value ($t = 1.71$) significant at the .05 level, in the direction of increased absenteeism. In general, the data gathered from the comparison groups showed no significant changes during matched baseline and intervention periods. In the one group where a significant change was shown, the change was actually in the direction of increased absenteeism.

DISCUSSION

The results of this study support the hypothesis that behavior modification can be utilized in an industrial setting to reduce absenteeism to a statistically significant degree.

Over a four-month period an 18.27% reduction in the absenteeism rate of 215 unionized employees was achieved. Over the same four-month period, a 13.79% average increase took place in the absenteeism rate of comparison groups. Note that the baseline rate of the experimental group was significantly lower than any of the comparison groups to start with. Further, the results showed that stretching the schedule of reinforcement from weekly to biweekly could be achieved without bringing about a significant increase in the rate of absenteeism. Also, since the program was designed so that an employee had to be on time before an incentive was given, a decrease in tardiness was also observed.

The incentive system was terminated after 16 weeks. The reason for doing so was that a bargaining date for a new union contract was approaching and the company did not want to find itself negotiating this incentive system into the contract. Thus, while the results of this study are encouraging, further work is necessary relative to the long-range maintenance of such behavior.

It would be interesting to compare the present study with a previous study by Lawler and Hackman (1969). They also decreased absenteeism by using an incentive system. However, Lawler and Hackman (1969) emphasized participation as the explanatory variable. In this study, emphasis was placed on behavioral principles as bringing about the decrease in absenteeism. Does this suggest two ways of decreasing absenteeism in industrial organizations—participation and behavior modification?

Gamboa (1974) suggested that processes given the label "participation" can simply be much better understood and operationalized using behavior modification principles. Thus, participation and behavior modification are not two ways of decreasing absenteeism (or improving any other behavior). Rather, behavior modification simply helps to make the concept of participation operative and effective. It is the consideration of these issues, both theoretically and empirically, which promises a continued growing interest in the application of behavior modification in industrial settings in the years to come.

REFERENCES

Adam, E. E. An analysis of changes in performance quality with operant conditioning procedures. *Journal of Applied Psychology,* 1972, **56**, 480–486.

Brethower, K. S. Performance indicators as feedback in business settings. Unpublished doctoral dissertation, University of Michigan, 1973.

Cherrington, D. J., Reitz, H. J., & Scott, W. E. Effects of contingent and noncontingent reward on the relationship between satisfaction and task performance. *Journal of Applied Psychology,* 1971, **55**, 531–536.

Conversation with B. F. Skinner. *Organizational Dynamics,* 1973, **1**(3), 31–40.

Gamboa, V. Beyond Skinner with dignity: An investigation of the application of behavior modification in industrial settings. Unpublished doctoral dissertation, University of Michigan, 1974.

Jablonsky, S. F., & DeVries, D. L. Operant conditioning principles extrapolated to the theory of management. *Organizational Behavior and Human Performance,* 1972, 7, 340–358.

Lawler, E. F., & Hackman, J. R. Impact of employee participation in the development of pay incentive plans: A field experiment. *Journal of Applied Psychology,* 1969, 53, 467–471.

Mann, F., & Baumgartel, H. *Absences and employee attitudes in an electric power company.* Ann Arbor, Mich.: University of Michigan, Institute for Social Research, 1952.

Mann, F. C., & Sparling, J. E. Changing absence rates: An application of research findings. *Personnel,* 1956, **32**, 392–408.

New tool: "Reinforcement" for good work. *Business Week,* December 18, 1971, p. 76.

Nord, W. R. Beyond the teaching machine: The neglected area of operant conditioning in the theory and practice of management. In W. R. Nord (Ed.), *Concepts and controversy in organizational behavior.* Pacific Palisades, Calif.: Goodyear, 1972, 145–168.

Porter, L. W. Turning work into non-work: The rewarding environment. In M. D. Dunnette (Ed.), *Work and non-work in the year 2001.* Monterey, Calif.: Brooks/Cole, 1973.

Where Skinner's theories work. *Business Week,* December 2, 1972, p. 64.

Yukl, G., Wexley, K. N., & Seymore, J. E. Effectiveness of pay incentives under variable ratio and continuous reinforcement schedules. *Journal of Applied Psychology,* 1972, **56**, 19–23.

(Received November 12, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 6, 699–704

USE OF JOB INFORMATION AS A CRITERION IN EMPLOYMENT DECISIONS OF INTERVIEWERS

YOASH WIENER [1]
Case Western Reserve University

MARK L. SCHNEIDERMAN
Personnel Research and Development Corporation, Cleveland, Ohio

Two factorially designed experiments investigated the extent to which job information was used by interviewers as a criterion for their decisions. In the first experiment subjects were experienced ($n = 48$) and inexperienced ($n = 48$) interviewers. All subjects received identical job information. Applicant information was defined as either relevant or irrelevant to the job information. Subjects' judgments were based on composites of relevant and irrelevant information segments. Each of the segments was either favorable or unfavorable. The second experiment replicated the first one, except that subjects ($N = 48$) did not receive job information. The main findings indicated extensive use of job information. Availability of such information reduced the effect of irrelevant attributes on decisions but did not eliminate it. Use of job information was not enhanced by experience in interviewing.

A factor which may contribute to biases in hiring decisions by interviewers is a reliance on personal hunches, theories, and stereotypes as a criterion for weighing and evaluating applicant information. The McGill University studies (Webster, 1964) strongly suggest that interviewers make hiring decisions by matching and comparing job applicants against a stereotype of an ideal applicant. These findings indicate that if the match is close the applicant is more likely to be hired; otherwise, he is more likely to be rejected. Consequently, the degree of the validity of the decision is a function of the fit of the stereotype with actual job requirements.

One way to reduce the potential bias involved in such personal stereotypes is to make available to interviewers a set of objective job requirements to be employed as criteria for evaluation of applicant information. The more interviewers use such job information, the less they are likely to rely on personal stereotypes.

Only one study to date (Langdale & Weitz, 1973) has investigated the availability of job information as a factor in employment decisions. In that study, two groups of personnel interviewers were given identical applicant information for evaluation. One group was given very little job information, while the other group received a rather full description of the job. The major finding was that interviewer agreement was far superior for the group having more complete job information.

The present article reports two experiments investigating the extent to which interviewers tend to use available job information as a basis for employment decisions and the possible conditions that may enhance or decrease such use. When an interviewer possesses job information, any applicant information he collects can be classified as either relevant or irrelevant to the available job information. Job information is highly used when the interviewer relies mostly on relevant applicant information and neglects the irrelevant information in his decision. Experiment 1 ascertains the relative contribution of relevant and irrelevant applicant information to the hiring decision when job information is available to interviewers. Experiment 2 ascertains the relative contribution of these information segments when job information is not available.

EXPERIMENT 1

Method

Subjects

Ninety-six subjects of whom half were experienced interviewers and the other half were inexperienced interviewers participated in the experiment. The experienced interviewers were drawn from local industrial companies and consulting firms. Each of them had been interviewing job applicants on a regular basis for at least six months and had been responsible for making either employment decisions or recommendations. The inexperienced interviewers were drawn from low- and middle-management personnel in local companies where they had not had any responsibility in making employment decisions or interviewing.

---

[1] Requests for reprints should be sent to Yoash Wiener, Department of Psychology, Case Western Reserve University, Cleveland, Ohio 44106.

Materials

Job information. An identical set of data about the job of a staff engineer was given to all subjects. This particular job was selected because the subjects, both experienced and inexperienced interviewers, were assumed to have the same degree of general familiarity with it. The job was specific enough to ensure that even the experienced subjects had not actually interviewed applicants for that position. Thus, it was possible to avoid the problem found by Webster (1964) wherein experienced interviewers may be locked into their own company's values, regardless of the given job information.

The job information presented to the subjects was adopted from actual job analysis data of the position staff engineer in a large local manufacturing company. It consisted of a list of specifications of the personal requirements of a staff engineer job for that company. An example of one item from this list of job requirements is:

> Supervision and work with others: The applicant must be able to work independently and with minimal supervision. Normally, he would be given a general assignment and the responsibility to follow it through to conclusion. Primarily, he would work alone. He does not supervise anybody and his relations with others are limited to exchanging information.

In order to approximate a realistic format of a complete set of job information data, subjects were also presented with a description of job duties and responsibilities.

Applicant information. This information was presented in a written format of 20 descriptive statements characterizing each applicant. Similar procedures have been commonly used in interview research (e.g., Carlson, 1967; Hollman, 1972). The descriptive statements used were selected from a pool of 730 items (Hakel & Dunnette, 1970) that had previously been judged on the basis of favorability, frequency of occurrence and importance to specific jobs (including the job, engineer). The process of item selection was conducted in four steps: (*a*) All items in the category engineer were evaluated by two judges as either relevant or irrelevant to the job specifications. The relevancy judgment was a straightforward procedure. For example, the job specifications excluded the need for effective interpersonal relations. Thus, any statement relating to interpersonal skills was irrelevant. (*b*) The relevant and irrelevant groups of items were further divided into favorable and unfavorable categories according to the ratings of Hakel and Dunnette (1970). Statements with a mean rating of 5.00 and above were labeled favorable and those with a mean rating of 3.00 or below were labeled unfavorable. The rest of the items were discarded. (*c*) The remaining items were arranged into opposing pairs on the basis of their favorability. In each of the relevant and irrelevant groups of items, favorable statements were matched with their unfavorable counterparts. For example, the favorable relevant item "is a college graduate" was paired with its unfavorable relevant counterpart "has had two years of college." (*d*) Twenty pairs were selected out of the pool of matched items, 10 relevant and 10 irrelevant pairs. Each pair contained one favorable item and one unfavorable item.

The items were arranged to form four applicant information sets. Each set consisted of 20 statements, half of which were relevant and half of which were irrelevant. The relevant and irrelevant segments were either favorable or unfavorable. To keep the task as realistic as possible the information did not consist of purely favorable or unfavorable statements. That is, the favorable segment consisted of 7 favorable and 3 unfavorable statements; vice versa for the unfavorable segment of the information. A pilot study showed that the four applicant information sets were perceived by subjects as realistic and credible. The 20 statements in each of the applicant information conditions were arranged in three different random orders. Each subject, however, was presented with only one set of information.

Design and Procedure

The four applicant information conditions were examined in a factorial design: (*a*) favorable relevant–favorable irrelevant; (*b*) favorable relevant–unfavorable irrelevant; (*c*) unfavorable relevant–favorable irrelevant; and (*d*) unfavorable relevant–unfavorable irrelevant. Subjects were assigned randomly to these conditions, each subject belonging to only one condition. This design cannot provide data about the main effect of relevant versus irrelevant information per se, since each subject received both relevant and irrelevant information and his employment decision was based on the composite. However, the design can provide data about the main effects of two independent variables: favorability versus unfavorability of relevant information and favorability versus unfavorability of irrelevant information. These main effects indicate the impact of each of these information types on the decision. The more interviewers rely on a given information segment—either relevant or irrelevant—in their decisions, the larger the difference between the favorable and unfavorable conditions of this information segment. That is, the larger the main effect. Omega-square values obtained from these main effects enable a separate assessment and comparison of the percentage of decision variance accounted for by the relevant and irrelevant applicant information segments or the relative contribution of each type of information to the hiring decision. Consequently, an inference about the extent of job information use is possible. An assessment of the interaction effect of the two favorability variables

TABLE 1

EMPLOYMENT DECISIONS AS A FUNCTION OF FAVORABILITY OF RELEVANT AND IRRELEVANT INFORMATION FOR EXPERIENCED AND INEXPERIENCED INTERVIEWERS

Variable	Experiment 1				Experiment 2	
	Experienced		Inexperienced		Inexperienced	
	Favorable irrelevant	Unfavorable irrelevant	Favorable irrelevant	Unfavorable irrelevant	Favorable irrelevant	Unfavorable irrelevant
Favorable relevant						
M	7.58[a]	6.25	9.33	8.25	7.83	4.92
SD	2.81	2.01	2.10	2.09	2.61	2.53
Unfavorable relevant						
M	3.50	1.75	6.08	2.25	5.33	2.25
SD	1.80	1.36	3.17	1.36	1.75	2.00

Note. n = 12 for each group.

[a] Mean value on the degree of rejection–acceptance scale. The scale ranged from very strongly reject (1) to very strongly accept (12).

may ascertain whether the extent to which irrelevant information affects the decision depends on the favorability level of the relevant information. The design also allows the determination of the degree to which experience in interviewing, a third independent variable, affects reliance on relevant and irrelevant applicant information, and consequently an inference about the effect of such experience on use of job information.

Each subject was presented with a booklet that included instructions, job information, applicant information, and the rating scales, in that order. Subjects were instructed "to assume that they are in a capacity to make an employment decision for a hypothetical company" and to make the decision on the basis of the information supplied. No restrictions were made on how to evaluate the information. No time limit was imposed.

Dependent measures. The *interview decision* was measured in two steps. The first was a dichotomous decision of either accept or reject. This approximates the type of decision generally made in a real-life employment interview. The second step was to request subjects to respond to a 6-point scale indicating the degree of strength of the acceptance or rejection decision. A subject who first made the decision of accept responded to a degree of acceptance scale, whereas a subject who initially made a rejection decision responded to a degree of rejection scale. For the purpose of analysis, these two scales were combined, so each subject received a decision score in the range of 1–12, indicating a decision continuum from very strongly reject to very strongly accept. This scaling procedure did not allow subjects to remain neutral or not make a decision. The other dependent variable—the *degree of confidence* subjects placed in their decision—was measured by a 7-point bipolar scale.

Results and Discussion

Decision Findings

Table 1 (Experiment 1) shows mean scores of the strength of accept–reject decision of the eight treatment groups ($n = 12$ for each group).

Table 2 (Experiment 1) summarizes the analysis of variance of the decision data. The main effects of favorability of relevant information and favorability of irrelevant information were significant. They accounted for decision variance proportions of 42% and 8%, respectively, as measured by omega–square values. These values indicate a considerable reliance on relevant applicant information when job information was available to interviewers. However, availability of such information did not completely eliminate the effect of irrelevant applicant information. It was much smaller than the effect of the relevant information, but still contributed significantly to the decision. This pattern of effects strongly indicates that the job information data were used by the interviewers as a criterion for their decision.

There was a tendency of a Favorability of Relevant Information × Favorability of Irrelevant Information interaction ($p < .10$). The pattern of means in Table 1 (Experiment 1) indicates that the favorability of irrelevant information had more effect on subjects' decisions when the relevant information was unfavorable than when the relevant information was favorable. Previous research found that interviewers are more influenced by unfavorable than by favorable information (Bolster & Springbett, 1961). Despite this predisposition, the present findings suggest that interviewers tend to pay more attention and give more weight to irrelevant information when faced with unfavorable relevant information than when faced with favorable relevant information. That is, when interviewers possess unfavorable relevant information they seem to seek additional available information for making a decision. When they possess favorable relevant information they tend less to seek additional information for their decisions.

The main effect of interviewers' experience was significant. The pattern of the means (Table 1, Experiment 1) shows that experienced interviewers were more rejecting than the inexperienced ones. Similar results were obtained in previous studies (Springbett, 1954; Webster, 1964). Possibly, because of organizational reward systems, experienced interviewers are more cautious in making their decisions. Interviewers' experience did not interact with any of the applicant information variables. This shows that the experienced interviewers did not differ from the inexperienced ones in the relative amount of relevant and irrelevant information they used in making their decisions. Experience does not seem to affect use of job information. Langdale and Weitz (1973) also found that experience did not affect interviewer agreement when job information was available.

Confidence Findings

A 2 × 2 × 2 analysis of variance on scores of degree of confidence in the hiring decision showed a significant main effect only for favorability of relevant information ($F = 5.03$, $df = 1/88$, $p < .05$). As indicated in Table 3 (Experiment 1), the more unfavorable the relevant applicant information, the larger the confidence interviewers have in their decisions. The main effect of favorability of irrelevant information was not significant, but the same pattern of means as for the relevant information was found. Interviewers seem to have slightly greater confidence in their decisions when they are based on unfavorable information. Additional support for this pattern comes from the −.35 correlation between the confidence ratings and the accept–reject scores. Interviewers are more confident in their decision when they reject an applicant.

TABLE 2

ANALYSIS OF VARIANCE FOR DECISION SCORES

Source	df	MS	F	omega²
Experiment 1				
Favorability of relevant information (A)	1	477.04	92.77**	.42
Favorability of irrelevant information (B)	1	96.00	18.67**	.08
Experience (C)	1	70.04	13.62**	.06
A × B	1	15.04	2.93*	.01
A × C	1	.67	.13	
B × C	1	5.04	.98	
A × B × C	1	8.17	1.59	
Within error	88	5.14		
Experiment 2				
Favorability of relevant information (A)	1	80.09	14.48**	.17
Favorability of irrelevant information (B)	1	108.00	19.53**	.24
A × B	1	.08	.01	
Within error	44	5.53		

* $p < .10$.
** $p < .01$.

The main effect of experience was not significant. Also, experience did not interact with any of the information variables so as to moderate their effect on decision confidence. The higher degree of confidence when decision and information are negative may give additional support to the contention that interviewers attach more weight to unfavorable information. Decisions based on information that is highly weighed produce a higher degree of confidence.

There was a significant Favorability of Relevant Information × Favorability of Irrelevant Information interaction ($F = 16.90$, $df = 1/88$, $p < .01$). The pattern of the means in Table 3 (Experiment 1) indicates that decision confidence was affected, in addition to the favorability of applicant information, also by the internal consistency of this information. Subjects had most confidence when the information was negative and consistent (unfavorable relevant–unfavorable irrelevant condition). The second highest degree of decision confidence was exhibited when the information was positive and consistent. When the information was inconsistent, that is, the relevant segment was favorable and the irrelevant segment was unfavorable, or vice versa, confidence was the lowest.

EXPERIMENT 2

Method

Subjects

All subjects were inexperienced interviewers. There were 48 subjects, drawn from the same population of the inexperienced subjects in Experiment 1.

Procedure

Experiment 2 was identical to Experiment 1, except the booklet presented to subjects did not include job information.

Design

Since the variable of experience was omitted in this experiment, its design was a 2 × 2 factorial. The relative contribution of the relevant and irrelevant applicant information segments to the hiring decision was ascertained and compared with Experiment 1. This comparison enabled the determination of the extent to which the effects of the relevant and irrelevant information in Experiment 1 differed from the effects of the same information segments when job information was not available to interviewers. The magnitude of this difference indicates the impact of job information on applicant evaluation and employment decisions.

Results

Decision Findings

The 2 × 2 analysis of variance of the decision data is presented in Table 2 (Experiment 2). The main effects of favorability of relevant and irrelevant information were significant. They accounted for decision variance proportions of 17% and 24%, respectively. A comparison with the corresponding proportions in Experiment 1 (42% and 8%) indicates that job information was used to a considerable extent by interviewers as a criterion for hiring decisions. Its availability considerably reduced the effect of irrelevant information and enhanced the effect of relevant information on hiring decisions.

TABLE 3

CONFIDENCE IN DECISION AS A FUNCTION OF FAVORABILITY OF RELEVANT AND IRRELEVANT INFORMATION

Variable	Experiment 1[a]		Experiment 2	
	Favorable irrelevant	Unfavorable irrelevant	Favorable irrelevant	Unfavorable irrelevant
Favorable relevant				
M	5.33[b]	4.38	4.92	5.66
SD	1.58	1.38	1.75	1.10
Unfavorable relevant				
M	4.67	6.50	5.08	6.00
SD	1.72	1.14	1.63	1.11

[a] The two levels of experience were pooled together since no differences were found between them.

[b] Mean confidence scores are on a 7-point bipolar confidence in decision scale with lower values indicating less confidence.

Availability of job information affected also the harshness of the hiring decisions. The overall decision mean for the subjects in Experiment 2 was 5.08. The corresponding mean for the inexperienced interviewers in Experiment 1 was 6.48 (Table 1, Experiments 1 and 2). Interviewers, then, seem to make more negative decisions when job information is not available for them. Experiment 1 showed that interviewers felt more comfortable when they made negative decisions, possibly because they perceived that it reduced the risk of being blamed for a bad decision. It is quite possible that making decisions on the basis of job information decreases such an apprehension and facilitates positive decisions.

Confidence Findings

Table 3 (Experiment 2) presents the means of decision confidence scores for all information conditions. Analysis of variance did not reveal any significant effect, but the main effect of favorability of irrelevant information approached significance ($F = 3.72$, $df = 1/44$, $p < .10$). As in Experiment 1, interviewers tended to have more confidence in their decisions when they were based on unfavorable information, but in Experiment 1 this effect was more pronounced for the relevant information. The reason for this difference possibly lies in the different degree of interviewer reliance on relevant and irrelevant information in the two experiments. In Experiment 1 they relied mostly on relevant information and in Experiment 2 they relied more on irrelevant information.

There was no significant interaction of Favorability of Relevant Information × Favorability of Irrelevant Information, as was found in Experiment 1. In Experiment 2, then, consistency of applicant information did not emerge as a factor affecting decision confidence. It is possible that the availability of job information as a criterion for a hiring decision makes the consistency of applicant information more salient for interviewers. Thus, the same sets of applicant information may have been perceived differently in the two experiments: as more inconsistent in Experiment 1 and less inconsistent in Experiment 2.

GENERAL CONCLUSIONS

The study shows that job information was indeed extensively used when available to interviewers. The contribution of relevant applicant information to the decision variance increased from 17% to 42%, whereas the contribution of irrelevant information decreased from 24% to 8%. The decrease in the variance proportion of the irrelevant information indicates a decline in the effect of personal subjective stereotypes on the hiring decision. The fact that the effect of irrelevant information had not been completely eliminated when job information was available, points at the strong predisposition of interviewers to hold personal stereotypes of an ideal applicant.

The effect of job information in the less structured and controlled situations of real-life decision making may be less dramatic than the effect observed in the present experiments. However, the conclusion remains that job information should be made readily available to interviewers. Since interviewers seem strongly inclined to use job information, care should be taken that this information is accurate. Otherwise, the availability of such information would hurt the quality of employment decisions. Rigorous procedures of job analysis can ensure sufficient accuracy of job information.

When job information was available to interviewers they tended to make more positive decisions. Further research is needed to find out whether availability of job information restores more balance to the relative weight interviewers give to favorable and unfavorable information, and whether, indeed, it produces more valid decisions.

Experience in interviewing did not seem to enhance a more extensive use of job information. This is not very surprising, since experienced interviewers probably tend to be locked into their companies' value systems, or into their own personal stereotypes, more than inexperienced interviewers are (Webster, 1964). It is possible that some training in the use of job information will benefit both experienced and inexperienced interviewers. In general, experienced interviewers tend to make more negative decisions than inexperienced ones.

Confidence in the employment decision was not found to be related to availability of job information. It was found to be affected by the favorability of applicant information and the internal consistency of the information. The more unfavorable the applicant information and the more consistent it is, the higher is the confidence in the decision. The predisposition of interviewers to feel more comfortable with unfavorable applicant information and with negative decisions is probably a result of the imbalance in weights given by interviewers to favorable and unfavorable applicant information.

REFERENCES

Bolster, B. I., & Springbett, B. M. The reaction of interviewers to favorable and unfavorable information. *Journal of Applied Psychology*, 1961, **45**, 97–103.

Carlson, R. E. Selection interviewer decision: The relative influence of appearance on factual written information on an interviewer's final rating. *Journal of Applied Psychology*, 1967, **51**, 461–468.

Hakel, M. D., & Dunnette, M. D. Checklists for describing job applicants. *Industrial Relations Center Bulletin*, **51**, March 1970.

Hollman, T. D. Employment interviewers' errors in processing positive and negative information. *Journal of Applied Psychology*, 1972, **56**, 130–134.

Langdale, J. A., & Weitz, J. Estimating the influence of job information on interviewer agreement. *Journal of Applied Psychology*, 1973, **57**, 23–27.

Springbett, B. M. Series effects in the employment interview. Unpublished doctoral dissertation, McGill University, 1954.

Webster, E. C. (Ed.), *Decision making in the employment interview*. Montreal: Industrial Relations Center, McGill University, 1964.

(Received February 4, 1974)

Journal of Applied Psychology
1974, Vol. 59, No. 6, 705–711

RACE AND SEX AS DETERMINANTS OF RATINGS BY POTENTIAL EMPLOYERS IN A SIMULATED WORK-SAMPLING TASK[1]

W. CLAY HAMNER,[2] JAY S. KIM,[3] LLOYD BAIRD[4]
AND WILLIAM J. BIGONESS[5]

Graduate School of Business, Michigan State University

This investigation examined the way the sex and race of the rater and the sex and race of the ratee influence assessments of ratee performance on a simulated work-sampling task. Undergraduate students ($n = 36$) were asked to assume the role of a manager and rate all eight combinations of male–female and black–white performers. Results indicated that sex–race stereotypes do influence assessments of behavior on a work-sampling task even when objective measures are defined. Findings are discussed in terms of potential methods for overcoming these biases in order to use more effectively the work samples for selection and promotion decisions.

A problem of current concern to the personnel psychologist is race and sex group discrimination in employment hiring. Numerous articles have appeared discussing methodological considerations relevant to discrimination in using employment tests as a method of employee selection (e.g., see Einhorn & Bass, 1971). Invariably, definitions and discussions of employment test bias or unfair discrimination refer to the important relationship between tests and job performance. The assumption in most of these reports is that the criterion of job performance is acceptable and accurate (Gordon, 1970).

Boehm (1972) and Einhorn and Bass (1971) have pointed out that even if tests are used appropriately, discrimination in selection can still occur if the criterion measures themselves are "biased" or "unfair" with respect to different subgroups. Generally the measure of job performance used is a subjective performance appraisal score, and therefore a major question still remains to be answered by industrial psychologists: Are worker assessments free from rater bias?

Kavanagh (1971) recommended that a rating of a specific stimulus is reliable to the extent that independent raters agree, or alternatively, that consistency over time can be demonstrated. While the argument for performance rating consistency is necessary for establishing valid performance criteria, the presence of high interrater reliability unfortunately does not preclude the presence of bias in the performance appraisal. First, interrater reliability scores do not reflect whether or not the ratings are correlated with objective measures of job performance. Second, and more important, since much of the work on interrater reliability has been done using homogeneous (i.e., white male) groups, the presence or absence of rater bias is not accurately measured. Along these same lines, if multiple raters (e.g., a supervisor and his boss) both share the same bias, then a high interrater reliability score could be possible, even when bias is present. Flaugher, Campbell, and Pike (1969), drawing on research conducted by the Educational Testing Service, concluded that if the chosen criterion itself is infused with the type of bias or unfairness that is being investigated in the predictor tests, the most elaborate of statistical manipulations is unlikely to detect its presence in the predictor.

Gael and Grant (1972) found that the quality of the criterion often determined the results of studies testing for differential validity. As a general rule, most studies showing a lack of differential validity have used better than average objective criterion measures, while most of the studies supporting differential validity rely on subjective, poorly determined criteria. Based on these findings, it seems reasonable to suspect that the problem is not just with the selection tests but also with the fact that the majority criterion performance is higher than the minority group performance (see Schmidt & Hunter, 1974) when subjective rating scales are used. Therefore, it seems imperative that industrial psychologists examine such performance ratings for bias as well as consistency.

---

[1] This research was supported by a grant to the first author from the Graduate School of Business Administration, Michigan State University.

[2] Requests for reprints should be sent to W. Clay Hamner, Room 239 Eppley Center, Michigan State University, East Lansing, Michigan 48824.

[3] Now at the Ohio State University.

[4] Now at Boston University.

[5] Now at the University of Maine.

While the courts have not yet examined in detail the criterion problem as it relates to selection bias, several recent statements by the courts indicate that they are concerned that validation of selection tests alone is not sufficient for the elimination of hiring and promotion biases. For example, Judge Brown, speaking for a unanimous court in Rowe versus General Motors Corporation,[6] stated that personnel decisions which depend almost entirely on the subjective evaluation of the immediate supervisor are a ready method of discrimination against blacks because bias can be covertly concealed with little chance of detection.

To overcome the subjective nature of the criterion ratings, many knowledgeable industrial psychologists have suggested that a job simulation test (i.e., work sampling) should be made a part of the selection procedure. This method of selection, according to Wernimont and Campbell (1968) and O'Leary (1973) should reduce discrimination in the selection process by making the selection criteria relevant and by removing unfairness of content found in many selection tests.

The purpose of this paper is to examine two questions which have been raised in the above discussion. First, are performance ratings free of rater bias when objective measures have been defined for the raters? This question has been raised by the findings of Boehm (1972), Bass and Turner (1973), Flaugher et al. (1969), and Gael and Grant (1972).

Second, are the assessments of work sampling unbiased, as hypothesized by Wernimont and Campbell (1968) and O'Leary (1973)? While some work-sampling tests are "prescored" (e.g., Thurstone Typing Test) and therefore should be an accurate measure of work-sampling performance, other work-sampling tests (e.g., see Campion, 1972) require evaluations by "job experts" in the rating process. It is possible that these judgments are open to bias, just as are other supervisory rating techniques. This is especially true in the assessment center setting. The American Telephone and Telegraph studies (Bray & Campbell, 1968) and the in-basket studies (Lopez, 1965) are examples of simulated work-sampling exercises which used ratings of behavior as measures of the performance level achieved.

The present study employed a controlled laboratory setting to examine what effect the sex and race of the rater and the sex and race of the job applicants had on the ratings in a simulated work-sampling task.

METHOD

Subjects and Design

Thirty-six undergraduate business students volunteered to serve as subjects in this laboratory experiment. A 2 × 2 × 2 × 2 × 2 factorial design was used to test the effects of rater's race (black versus white), rater's sex (male versus female), ratee's race (black versus white), ratee's sex (male versus female), and ratee's level of performance (high versus low) on assessments of the ratee's overall performance on the task. Rater's race and sex were subject classification factors, while ratee's race, sex, and level of performance were repeated factors. The subjects included 9 black males, 9 black females, 9 white males, and 9 white females. Each subject rated the ratees, four high performers and four low performers. The four race–sex combinations, that is, black male, black female, white male, and white female were represented in both of the two performance categories.

The subjects were assigned the role of a grocery store manager and potential employer of the eight ratees, all of whom had supposedly applied for a job as a stock worker in that grocery store. As part of the screening process, the ratees were asked to complete a three-minute work-sampling task. The task involved removing large Number 10 cans from an opened case and placing them on the grocery shelves provided. Based on the personal experience of the researchers and on a time study, it was determined that the average worker could stock 36 cans in a three-minute period. The high-performance ratees in this experiment stocked 48 cans during the three-minute period, while the low-performance ratees stocked 24 cans during this same period. The ratees were trained at the task so that certain dimensions, motions (use of their hands, etc.), facial expressions, mode of dress, and quality of work (labels faced outward, etc.) were essentially identical. For example, each ratee (both male and female) wore dark slacks, a light shirt, and a dark grocer's apron. Also, each ratee used both hands while stacking the large cans. A panel of graduate students chosen to act as judges agreed that the ratees did in fact appear essentially identical on these dimensions. Therefore, the only variance of significance among the ratees would be along the race, sex, or level of performance (quantity only) dimensions.

---

[6] Rowe v. General Motors Corporation, 457 F. 2d 348 (5th Cir. 1972).

The Michigan State University Instructional Media Center designed a "grocery store" set and filmed the performance of each of the eight ratees. The set was realistic and resembled the stock section of a large grocery chain. The film was then available to be shown on closed-circuit television to the 36 subjects. The order of presentation of the ratees on the film was as follows: (*a*) low-performing black male, (*b*) high-performing white female, (*c*) low-performing black female, (*d*) high-performing white male, (*e*) low-performing white male, (*f*) high-performing black female, (*g*) high-performing black male, and (*h*) low-performing white female. Filming of the task performance was the media chosen so that consistency in performance could be maintained. Lewin, Dubno, and Akula (1971) found that face-to-face interaction is apparently not crucial in ratings. Ratings made after watching ratees on videotape were almost identical to those made after fairly extensive face-to-face interaction.

Procedure[7]

Data were collected from subjects who participated in approximately four equal size groups, although the subjects' ratings were done independently of each other. There were approximately an equal number of sex-race rater combinations at each showing.

Each subject was given an instruction sheet describing the purpose of the film. Also on the instruction sheet, the subjects were given eight large spaces in which they were asked to make comments about each applicant's performance to aid them in filling out the recommendation forms. The film was stopped at the end of each three-minute segment to allow the subjects to complete their comments. The specific instructions were as follows:

> As the manager of the *XYZ* grocery store, you anticipate several vacancies for stock personnel in the near future. You are considering 8 applicants for these positions. None of the applicants have any prior work experience, and all have a high school education. Your assistant manager has interviewed all applicants, and he screened out all but 8 from further consideration. He then had each of these 8 candidates complete a standard three minute "work sample" test. In this test, the applicants were asked to stock shelves for a three minute period of time. Based on time-and-motion studies, you have determined that an average worker, stocking large #10 cans, will be able to stock 6 cases (6 cans per case for a total of 36 cans) in a three minute period. Since you are very busy, you have asked the assistant manager to film the "work sample" results. You are now reviewing the film in order to make certain recommendations to your assistant manager concerning potential future employment of these workers. In order to aid you in filling out the recommendation forms which will be given to you after viewing the film, please make relevant notes about each applicant in the spaces provided below as you are viewing the film.

After the subjects had read the instructions, the investigator oriented the subjects by explaining that the number of openings was unclear at this time and that they should look for cues which would enable them to distinguish among the applicants if only three or four positions came open.

After viewing the film, each subject was given a packet containing eight ratee sheets, numbered consecutively from Worker 1 to Worker 8. Each worker was rated on his or her "overall task performance."[8]

Rating Scale

Ratings of the ratee's overall task performance obtained from the subjects at the end of the 24-minute film were the major dependent measure used. Rating Scale Content III format (see Parker, Taylor, Barrett, & Martens, 1959) was chosen, based on the fact that Taylor and his associates have shown that it is the best of four formats in terms of the problems of halo, leniency, reliability, and variability which generally surround the use of graphic rating scales. Each subject was asked to check the appropriate block on the scale that best described the applicant in question. The scale points ranged from 1 to 15, with 1 being the lowest possible rating on the scale. The rating categories were as follows: (*a*) weak in overall performance, failed to produce acceptable amount of production; (*b*) performance periodically fell below that which I would expect from similar individuals in this task; (*c*) did a satisfactory job, was an average worker in his performance; (*d*) was

---

[7] Postexperimental discussions with pilot studies' subjects and debriefing sessions with the subjects in this study revealed that the subjects were not aware of the purpose of this study (i.e., race and sex difference in performance ratings).

[8] In addition, the subjects rated the workers' task motivation and ability to perform the task, but the results of a multivariate analysis of variance used to analyze the data revealed that the findings for the three ratings were sufficiently similar that only the results for the overall performance ratings need be presented.

TABLE 1

CELL MEANS AND STANDARD DEVIATIONS OF THE RATINGS ON OVERALL TASK PERFORMANCE

Performance	Male rater				Female rater			
	Black rater		White rater		Black rater		White rater	
	Male applicant	Female applicant	Male applicant	Female applicant	Male applicant	Female applicant	Male applicant	Female applicant
Black applicant								
Low								
M	9.00	10.22	6.77	6.11	6.00	6.66	6.33	7.44
SD	4.26	2.16	2.94	2.47	2.23	3.55	3.50	2.87
High								
M	6.22	9.66	8.11	9.66	9.33	10.88	6.77	11.22
SD	2.99	4.12	1.83	3.00	2.50	2.73	2.56	4.03
White applicant								
Low								
M	4.88	6.44	5.44	5.77	5.00	5.22	6.55	5.77
SD	2.47	2.65	1.23	2.63	2.39	2.22	[illegible]	2.29
High								
M	9.55	9.77	10.11	10.88	9.22	11.66	10.00	11.66
SD	2.12	2.04	2.36	2.20	1.98	2.34	3.31	3.20

a good worker, his performance was better than average; and (*e*) an exceptionally good worker who will be equaled by few others.

RESULTS

Examination of the results of Table 1 indicates that the ratees' level of performance on the task corresponded to the raters' evaluation of this performance. High-performance ratees were rated as having significantly higher levels of overall performance than low-performance ratees (9.67 versus 6.48; $F = 70.45$, $df = 1/32$, $p < .001$). Omega-square values were computed to estimate the proportion of variance accounted for by level of performance of the ratees (Winer, 1971, p. 428). On this significant difference ($p < .001$), 30% of the variance was due to the main effect of level of performance. While this finding *alone* would indicate that evaluations of performance of the workers were free from rater bias, other findings render this conclusion invalid.

It was found that female ratees were rated higher than male ratees on their overall task performance (8.69 versus 7.46; $F = 16.79$, $df = 1/32$, $p < .001$; $\omega^2 = .08$). This was especially true for high-performing ratees. While low-performing male workers were rated only slightly lower than low-performing female workers (6.25 versus 6.71), high-performing male workers were rated markedly lower than high-performing female workers (8.66 versus 10.68; $F = 5.36$, $df = 1/32$, $p < .05$; $\omega^2 = .04$). It was also found that black raters tended to rate black ratees higher than white ratees (8.50 versus 7.72), while white raters tended to rate black ratees lower than white ratees (7.80 versus 8.27) on overall task performance ($F = 4.71$, $df = 1/32$, $p < .05$; $\omega^2 = .02$).

Another finding that is shown in the ratings on the overall task performance scale is that while high-performing blacks were rated only slightly higher than low-performing blacks, high-performing whites were rated significantly higher than low-performing whites. On this overall task performance scale, low-performing black workers and high-performing black workers were both rated as average workers (7.31 versus 8.98), while low-performing white workers were rated below average (5.63) and high-performing white workers were rated better than average (10.36). This significant interaction effect ($F = 19.56$, $df =$

1/32, $p < .001$; $\omega^2 = .09$) indicates a tendency to favor a high-performing white applicant over a high-performing black applicant and, at the same time, to favor a low-performing black applicant over a low-performing white applicant. This finding thus suggests that low-performing blacks are at an advantage and high-performing blacks at a disadvantage if assessments are used in the selection process.

While an examination of the omega-square values shows that the performance differences of the ratees accounted for 30% of the variance in the subjects' responses, an additional 23% of the variance in the subjects' responses was accounted for by various sex–race difference combinations. This finding suggests that even though performance differences were the major determinant of the subjects' ratings, race–sex differences also significantly affected the subjects' evaluations of the ratees.

DISCUSSION

To what extent are the assessments of work samplings biased against black and/or female applicants for a specific entry level job? Had we only considered the typical white male population of raters and ratees, we would have found that raters of high- versus low-performance workers were highly accurate in their assessment of overall task performance. High-performance applicants were rated significantly higher on this scale than were low-performance applicants. This finding would seem to support the contention made by Scott and Hamner (1973) that subjective ratings can be substituted for actual production records. Unfortunately, this may only be true for homogeneous worker populations.

When the race–sex combinations of the raters and ratees were considered, our data indicated that direct, systematic biases were present. First, in what is predominantly a male-dominated job, female applicants were seen as having a higher level of overall task performance than were male applicants for the same job. This may be due to the fact that this is a low-level, unskilled task which a male would be expected to hold, and when the female performs equally well on the task, she is seen as being a better performer than the male. This tendency to inflate rather than devalue the performance of women is contrary to previous findings (e.g., see Rosen & Jerdee, 1973). The main difference between the previous studies and the present one is that the present study used an unskilled task role, whereas the other studies used a managerial or professional task role. Therefore, it could be argued that females have a better chance of being hired for unskilled jobs, whereas males have a better chance of being hired for professional or managerial jobs.

A second systematic bias can be seen in the higher ratings that subjects gave to applicants of the same race. This finding supports the findings of Cox and Krumboltz (1958) and DeJung and Kaplan (1962). It also partially supports the findings by Flaugher et al. (1969) that blacks received distinctively higher ratings from supervisors of their own ethnic group than from white supervisors, and the findings by Farr, O'Leary, and Bartlett (1971) and Greenhaus and Gavin (1972) that white employees obtained higher supervisory ratings on overall job performance than did black employees.

As noted by Bass and Turner (1973), this finding alone would not necessarily prove that the criterion measure is unfairly discriminatory if there were actual differences in objective measures of job performance. Unfortunately, such is not the case. In the present study objective criteria were established, and the fact that blacks received significantly lower ratings than whites from white raters when performance levels were identical indicates a potentially serious problem of racial bias. In most industrial settings, the supervisors are predominantly white, and therefore, if the results of this laboratory study were supported by future investigations, it would mean that the use of performance evaluations as the criterion in job selection works to discriminate unfairly against black applicants.

The data indicate that the differential criterion bias is especially prevalent for the high-performing black applicant. We found that high-performing black ratees were rated only slightly higher than low-performing black ratees, while high-performing white ratees were rated significantly higher than low-performing white ratees. This finding tends to support the findings of Rotter and Rotter (1966) that raters generally gave higher ratings to black than to white ratees when performance was poor. Contrary to the conclusion drawn by Bass and Turner (1973), this finding seems to indicate that the subjects tended to rate the performance of white ratees more objectively than black ratees. Based on the last finding, it seems reasonable to conclude that some differential bias may be carried over to the job itself. For example, if salary increases and promotions are based on supervisory evaluations of employees, high-performing black employees would advance less rapidly than high-performing white employees.

These findings also indicate that the supervisory ratings as obtained would be inappropriate as criterion measures against which to validate tests, since females were rated differently than males and blacks were rated differently than whites, even though the performance of both groups was identical. To overcome this problem, training of the raters in the nature and use of rating scales to evaluate both periodic work performance, as recommended by Brown (1968) and Bass and Turner (1973), and work-sampling performance, as recommended by Campion (1972) and Bray and Campbell (1968), is necessary. Also, the raters could be exposed to training which emphasizes sex and race fairness to become aware of potential biases. Schmidt and Johnson (1973) have shown this to be an effective method of reducing racial bias among raters.

This is probably the first study to test the effect of the race and sex of the rater and the ratee on ratings given when an objective measure of performance was known. Care should be taken in attempting to generalize from this laboratory study with college students to organizational settings. Therefore, more research on the effect of race and sex bias on work-sampling assessments is indicated. A generalization from these results which does seem appropriate is that training in performance evaluation techniques *as well as* objectively defined criteria are needed before the employer will be able to hire the best qualified applicant for the job.

REFERENCES

Bass, A. R., & Turner, J. N. Ethnic group differences in relationships among criteria of job performance. *Journal of Applied Psychology*, 1973, **57**, 101-109.

Boehm, V. R. Negro-white differences in validity of employment and training selection procedures: Summary of research evidence. *Journal of Applied Psychology*, 1972, 56, 33-39.

Bray, D. W., & Campbell, R. J. Selection of salesmen by means of assessment centers. *Journal of Applied Psychology*, 1968, **52**, 36-41.

Brown, E. M. Influence of training, method, and relationship on the halo effect. *Journal of Applied Psychology*, 1968, **52**, 195-199.

Campion, J. E. Work sampling for personnel selection. *Journal of Applied Psychology*, 1972, **56**, 40-44.

Cox, J. A., & Krumboltz, J. D. Racial bias in peer ratings of basic airmen. *Sociometry*, 1958, **21**, 292-299.

DeJung, J. E., & Kaplan, H. Some differential effects of race of rater and ratee on early peer ratings of combat aptitude. *Journal of Applied Psychology*, 1962, **46**, 370-374.

Einhorn, H. J., & Bass, A. R. Methodological considerations relevant to discrimination in employment testing. *Psychological Bulletin*, 1971, **75**, 261-269.

Farr, J. L., O'Leary, S., & Bartlett, C. J. Ethnic group membership as a moderator of the prediction of job performance. *Personnel Psychology*, 1971, **24**, 609-636.

Flaugher, R. L., Campbell, J. T., & Pike, L. W. *Ethnic group membership as a moderator of supervisor's ratings.* (ETS Bulletin PR-69-5) Princeton, N.J.: Educational Testing Service, 1969.

Gael, S., & Grant, D. L. Employment test validation for minority and nonminority telephone company service representatives. *Journal of Applied Psychology*, 1972, **56**, 135-139.

Greenhaus, J. F., & Gavin, J. F. The relationship between expectancies and job behavior for white and black employees. *Personnel Psychology*, 1972, **25**, 449-455.

Gordon, M. E. The effect of the correctness of the behavior observed on the accuracy of ratings. *Organizational Behavior and Human Performance*, 1970, **5**, 366-379.

Kavanagh, M. J. The content issue in performance appraisal: A review. *Personnel Psychology*, 1971, **24**, 653-668.

Lewin, A. Y., Dubno, P., & Akula, W. G. Face to face interaction in the peer nomination process. *Journal of Applied Psychology*, 1971, **55**, 495-497.

Lopez, F. M. *Evaluating executive decision making: The in-basket technique.* New York: American Management Association, 1965.

O'Leary, L. R. Fair employment, sound psychometric practice, and reality: A dilemma and partial solution. *American Psychologist*, 1973, **28**, 147-150.

Parker, J. W., Taylor, E. K., Barrett, R. S., & Martens, L. Rating scale content: III. Relationship between supervisory and self-ratings. *Personnel Psychology*, 1959, **12**, 49–63.

Rosen, B., & Jerdee, T. H. The influence of sex-role stereotypes on evaluations of male and female supervisory behavior. *Journal of Applied Psychology*, 1973, **57**, 44–54.

Rotter, N., & Rotter, G. S. Race, work performance and merit rating: An experimental evaluation. Paper presented at the meeting of the Eastern Psychological Association, Philadelphia, April 1969.

Schmidt, F. L., & Hunter, J. E. Racial and ethnic bias in psychological tests: Divergent implications of two definitions of test bias. *American Psychologist*, 1974, **29**, 1–8.

Schmidt, F. L., & Johnson, R. H. The effect of race on peer ratings in an industrial situation. *Journal of Applied Psychology*, 1973, **57**, 237–241.

Scott, W. E., & Hamner, W. C. The influence of worker's performance patterns on the accuracy of supervisory ratings. Paper presented at the meeting of the Midwestern Psychological Association, Chicago, May 1973.

Wernimont, P. F., & Campbell, J. Signs, samples and criteria. *Journal of Applied Psychology*, 1968, **52**, 372–376.

Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1971.

(Received February 11, 1974)

Journal of Applied Psychology
1974, Vol. 59, No. 6, 712–716

EFFECTS OF OPPORTUNITY TO COMMUNICATE AND VISIBILITY OF INDIVIDUAL DECISIONS ON BEHAVIOR IN THE COMMON INTEREST[1]

THOMAS H. JERDEE AND BENSON ROSEN

Graduate School of Business Administration,
University of North Carolina at Chapel Hill

Two experiments examined the effects of opportunity to communicate and visibility of individual decisions on socially responsible behavior in group situations involving the "commons" problem. Two hundred and eighty college students, enrolled in a business administration course, were assigned roles as firm representatives and assembled in groups of 5 to bid on contracts under various conditions. Opportunity to communicate resulted in more socially responsible bidding, but visibility of individual decisions had no effect. Opportunity to communicate also helped to sustain socially responsible bidding in the face of a persistently defecting group member. Results are discussed in terms of implications for situations where socially responsible behavior is difficult to obtain because of the incentive structure.

Conflicts between the immediate interests of individuals or groups and the common interest of society have become increasingly prevalent in recent years. This problem has been called the "tragedy of the commons" by Hardin (1968) and Crowe (1969), who referred to difficulties that arose in England in the 19th century with regard to common pasturage. Each herdsman perceived a positive utility for increasing the size of his flock, but when all herdsmen acted on this perception, the result was serious depletion of pasturage.

More recently, Platt (1973) has discussed this problem in terms of reinforcement contingencies and analyzed various kinds of "social traps" in which the short-run gain of one individual, organization, or nation may involve a small, almost imperceptible cost to the common interest. In these situations, people tend to ignore the long-run common interest and to act in their own short-run interest. But when many people follow this course, the results for the common interest can be disastrous. These problems arise with regard to energy consumption, population control, pollution control, exchange rates, tariffs, fishing rights, and a host of other ecological and economic phenomena. They become more acute as resources become more scarce and as human interdependence increases. Consequently, there is an important need to understand the conditions in which people are more likely to behave in the common interest.

The formal properties of the problem of the commons have been analyzed by Kelley and Grzelak (1972). They described it as an *n*-person interdependency relationship, of which the Prisoner's Dilemma Game is the two-person case. They have demonstrated experimentally that the tendency to behave in a socially responsible manner in group situations is influenced by the nature of the payoff matrix. People are more likely to behave in the common interest when their short-run costs for doing so are smaller.

In the Kelley and Grzelak (1972) study and in another study by Kahan (1973), low rates of cooperative behavior were observed, and there were indications of a low level of awareness regarding the essential nature of the interdependency in the "commons" problem. This rather dismal state of affairs suggests that attention should be given to the means by which social awareness and commitment to behave in a socially responsible manner can be heightened. The problem is particularly challenging in situations where the long-run common interest is very important but where short-run payoffs cannot readily be altered in favor of the common good.

---

[1] The authors wish to thank Thomas Prestwich, Edward Bailey, Mary Mericle, and Sara Freedman for their assistance.

Requests for reprints may be sent to either author, Graduate School of Business Administration, University of North Carolina, Chapel Hill, North Carolina 27514.

An attractive possibility in these situations is the use of social sanctions. If individual members can be convinced that social acceptance and approval are dependent on behavior in the common interest, perhaps these contingencies can outweigh other more tangible short-run incentives for self-seeking behavior. Thus it may be useful to determine the conditions under which appropriate social sanctions are generated within a group. It seems likely that the visibility of individual members' behavior and the opportunity for group members to communicate freely with one another would both be important factors influencing socially responsible behavior.

Some support for the visibility factor is provided by findings of Schachter and Hall (1952) and Bennett (1955) that subjects were more willing to serve in experiments when they perceived a greater degree of commitment among their peers. Communication usually promotes cooperation in dyadic Prisoner's Dilemma Games but was found *not* to be effective in producing cooperative behavior in dyads in a "trucking game" situation (Deutsch & Krauss, 1962) and in triads in a coalition formation game (Vinacke, Mogy, Powers, Langan, & Beck, 1974). However, none of these studies involved a small-group commons situation, and the effects of visibility and communication on socially responsible behavior in small group settings involving the problem of the commons have not been established experimentally.

The present experiments were designed to investigate these effects in the context of a simulation of firms bidding on manufacturing contracts. Bids could be either "complete cost," involving a small short-run sacrifice for the long-run benefit of the society and the industry as a whole, or "cut cost," involving a short-run gain at the expense of society and the industry.

The first experiment examined the effects of free communication among group members and complete visibility of each member's decisions on the establishment of socially responsible complete-cost bidding behavior within the groups. The second experiment examined the power of free communication to *sustain* complete-cost bidding in the face of persistent defection on the part of one group member who was an accomplice of the experimenters.

EXPERIMENT 1

Method

Subjects and Procedure

Subjects were 180 undergraduate students enrolled in a business administration course at the University of North Carolina. Subjects were assembled in groups of 5 and seated around a table. The following written instructions were distributed and read aloud by the experimenter.

> Each of you in this room will be a representative of a different manufacturing firm. Your job will be to bid on contracts to manufacture a variety of goods. In each round of bidding, each firm will be awarded a contract, but depending upon how you bid and how other firms bid, your firm will make different amounts of profit. Your objective is to maximize the profits of your firm.
>
> On each round of bidding, your firm has a choice of submitting two different types of bids. The nature of these bids and their expected profits are described below.
>
> 1. *Complete Cost Bid.* This bid is relatively high because it takes into account the actual cost of production, using the finest materials available. In addition, this type of bid includes cost of indirect long-term expenses for anti-pollution measures, restoration of natural resources, and quality control devices. Since this bid is high, it might result in less profit, particularly if the other firms are not submitting complete cost bids. On the other hand, if you and most or all of the other firms submit complete cost bids, the cost of production will stabilize, the reputation of the industry will be enhanced, and every firm should experience higher profits.
>
> 2. *Cut Cost Bid.* This bid is relatively low since it covers only the expenses of the manufacturing process itself and ignores the additional expenses for pollution control, restoration of natural resources, and quality control. Because a cut cost bid is low it will probably lead to larger contracts and more profits. However, if you and most or all of the firms submit cut cost bids the reputation of the industry will decline. Accordingly, profits will be lower for you and for all of the other firms as well.
>
> You will be asked to bid on each of a series of contracts. Your choice involves making either a "complete cost" or a "cut cost" bid for each contract. The amount of profit that you make for your firm depends partially upon how you bid and

TABLE 1

PROFIT SCHEDULE FOR FIVE-FIRM INDUSTRY

Number of Firms Bidding Complete Cost	Profits of Each Complete-Cost Bidder	Number of Firms Bidding Cut Cost	Profits of Each Cut-Cost Bidder
0	—	5	$25
1	$20	4	$30
2	$25	3	$35
3	$30	2	$40
4	$35	1	$45
5	$40	0	—

> partially on how other firms bid. In any one bidding session, a "cut cost" bid will always be more profitable than a "complete cost" bid; however, the more firms bidding cut cost, the lower the general level of profitability in the industry. Therefore, on any one trial, your firm has a choice whether to bid "cut cost" so as to maximize profits, or to bid "complete cost" so as to raise the general level of profitability in the industry.
>
> You have been provided with a Profit Schedule showing the relationship between type of bid and amount of profit for a five-firm industry. Please examine this schedule carefully.
>
> Each firm will submit a bid for each contract by circling either "complete cost" or "cut cost" on the bidding form. The experimenter will collect the bids and calculate the profits for each firm on that particular contract.

The experiment consisted of a 2 × 2 factorial design, with nine groups of five subjects in each of the four cells. Manipulation of the independent variables (visibility of bids and opportunity to communicate) was accomplished through the following instructions.

> [*Low visibility*] The experimenter will then return the bidding form with your profit recorded next to your bid. Representatives will not be given information regarding the profits of other firms.
>
> [*High visibility*] The experimenter will then record on the blackboard each firm's profits.
>
> [*Low communication*] Please do not talk to any of the other participants during the remainder of the experiment. If you do, the control that is necessary for good research will be lost, and the experiment will be invalid.
>
> [*High communication*] Please do not talk to any of the other participants while bids are being made. After every round of bidding, there will be a one-minute period for open discussion among bidders.[2]

In order to emphasize the point that the objective was to maximize profits rather than to beat the other firms, a chart with various profit levels was posted on the wall, and the experimenter called attention to the goal of getting the highest possible total profits.

The profit schedule which was given to each subject is shown in Table 1. This schedule was thoroughly explained and illustrated by the experimenter. Subjects were not informed regarding the number of bidding trials, but in order to minimize terminal effects the bidding forms were numbered for 20 trials, and the experiment was run for only 10 trials. The dependent variable was the number of complete-cost bids on Trials 6–10.

Results and Discussion

Results of the first experiment are shown in Table 2. Communication had a pronounced effect. The mean number of complete-cost bids per trial on Trials 6–10 was approximately twice as high in the high-communication condition as in the low-communication condition ($F = 39.5$, $p < .01$). Subjects in the high-communication condition used the communication periods to discuss the need for socially responsible bidding, and in some instances they obtained verbal commitments to bid complete cost. In the low-communication condition this opportunity for persuasion and public commitment was not available, and subjects tended to bid cut cost.

Contrary to our expectations, visibility had no effect on bidding. This may be explained as follows: In the high-visibility-low-communication condition a cut-cost bidder, even though his action was visible, could continue his behavior with complete impunity, since the other bidders had no way to persuade him regarding the error of his ways. In the low-

TABLE 2

MEAN NUMBER OF COOPERATIVE BIDS (COMPLETE COST) PER TRIAL, EXPERIMENT 1

Trials	Low communication[a]		High communication[a]	
	Low visibility	High visibility	Low visibility	High visibility
1–5				
$\bar{X}$	1.9	2.6	3.0	2.8
SD	.8	1.1	.6	.9
6–10				
$\bar{X}$	1.6	1.9	3.7	3.6
SD	.5	.9	1.1	1.0

Note. All group *ns* = 5.
[a] Number of groups = 9 for low visibility and high visibility.

---

[2] In 6 of the 18 groups in the high-communication condition, communication was permitted on only every third trial. This restriction had a negligible effect.

visibility-high-communication condition, it apparently was enough for people to be convinced of the importance of complete-cost bidding and, in many instances, to have announced their intent to bid complete cost; individual visibility was not necessary to sustain socially responsible bidding when people could observe that the group as a whole was bidding complete cost. This suggests that if a norm of social responsibility is established, visibility of individual behavior may not be essential. Conversely, when a social responsibility norm does not exist, visibility of behavior may not have much effect unless it is accompanied by some kind of legal sanction.

In summary, results of the first experiment indicate the importance of communication as a means of establishing and maintaining socially responsible behavior in a commons situation. As long as subjects could discuss the problem with each other and determine that the group as a whole was committed to behave responsibly, they were able to "keep the faith."

But in many social situations there are likely to be a few defectors who, despite their promises to behave in the common interest, attempt to take advantage of others. Thus there is a question whether communication can sustain a norm of social responsibility in the face of persistent minority defection toward self-seeking behavior. This question was explored in the second experiment.

EXPERIMENT 2

In the second experiment we hypothesized that in the face of defection, socially responsible behavior could be maintained as long as there was continued opportunity to communicate but that social responsibility would deteriorate if no communication was possible.

This experiment employed a 2 × 2 design in which the independent variables were defection or cooperation of an accomplice and opportunity to communicate.

Method

Subjects and Procedures

Subjects were 100 undergraduate students enrolled in a business course at the University of North Carolina. Procedures were similar to those employed in the first experiment.

Communication was allowed after each of the first five trials, in order to build up a norm of social responsibility in all cells. Then, in the high-communication condition, communication was allowed after each subsequent trial as well, but in the low-communication condition no further communication was allowed. All subjects participated under low visibility. In the "defector" condition, an accomplice of the experimenters bid cut cost on every trial after the fifth trial, regardless of the communication condition. In the cooperator condition, an accomplice of the experimenters bid complete cost on every trial after the fifth trial. (Subjects were debriefed and informed of experimental results a few days after the experiment.)

Results and Discussion

As a check on the induction under full communication for the first five trials, the number of subjects in each cell bidding complete cost on the fifth trial was tabulated. The mean number of complete-cost bidders per group was 3.8 out of a possible 4, which indicated that communication had built up a norm of highly cooperative behavior by the fifth trial.

Treatment effects are shown in Table 3. Main effects were not significant for Trials 6–10 but had reached significance by the 10th trial (communication $F = 13.2$, $p < .01$; defector $F = 4.6$, $p < .05$). However, major interest is in the comparison between the two cells in which the accomplice defected on Trials 6–10. Here, communication had the predicted effect of sustaining cooperative bidding, while cooperation deteriorated in the low-communication cell ($t = 1.82$, $p < .03$).[3] In fact, on the 10th trial, after the defecting accomplice had bid cut cost four times in a row, the seven high-communication groups still produced an average of 3.1 cooperative bids versus only 1.3 for the six low-communication groups ($t = 2.86$, $p < .01$).

The results of these two experiments provide strong support for the importance of the opportunity to communicate in social situations involving the problem of the commons. In both experiments, communication was not imposed or required; the opportunity was simply provided to be used in any manner the subjects chose. In Experiment 1, communica-

---

[3] A one-tailed test was used because the direction of the effect was predicted on the basis of the strong effects of communication in Experiment 1.

TABLE 3

MEAN NUMBER OF COOPERATIVE BIDS (COMPLETE COST) PER TRIAL, EXPERIMENT 2

Trials	Low communication^a		High communication^b	
	Defecting accomplice	Cooperating accomplice	Defecting accomplice	Cooperating accomplice
1-5				
$\bar{X}$	3.3	3.4	3.2	3.2
SD	.3	.4	.6	.4
6-10				
$\bar{X}$	2.2	2.9	3.1	3.3
SD	.7	1.1	1.0	.9
10				
$\bar{X}$	1.3	2.5	3.1	3.8
SD	1.0	1.4	1.2	.4

Note. All group *n*s = 4.

[a] Number of groups = 6 for defecting accomplice and cooperating accomplice.

[b] Number of groups = 7 for defecting accomplice and 6 for cooperating accomplice.

tion was the critical factor in the development of socially responsible norms. In Experiment 2, communication was essential to establish and maintain norms of social responsibility in the face of a defecting group member.

In a recent study, Kelley and Stahelski (1970) discussed an "assimilation" effect in two-person, mixed-motive interactions. Subjects with cooperative goals gradually assimilated to the position of their competitive opponents. A similar assimilation phenomenon was observed in the defector–low-communication condition in Experiment 2. Group members who initially bid in a cooperative manner became increasingly competitive when confronted with the competitive accomplice. On the other hand, in the defector–communication condition, cooperation was sustained because communication periods between trials provided group members with an opportunity to establish cooperative coalitions among themselves. Their unanimous or nearly unanimous pledges of cooperation signaled to each participant that if he would only make a small sacrifice on the next trial, he would most likely receive a higher payoff on succeeding trials. Communication also allowed subjects to signal to each other the availability of what Platt (1973) called "counter-reinforcers," in the form of social approval for sacrifice in the common interest.

A practical implication of these findings is to focus attention on the importance of obtaining expressions of commitment to act in a socially responsible manner as a means of producing behavior in the common interest. We would expect this strategy to be especially effective in situations involving small face-to-face groups, where the extent of commitment within the group could readily be established for all to perceive. Many committee and work group situations are of this nature. The strategy would be more difficult to implement in larger groups, where the extent of commitment could not so readily be established. Perhaps, in these situations, the "linking pin" concept (Likert, 1967) could be employed. Commitment could be obtained within small groups, and each group leader could convey this commitment to the next level in the organization. In addition, investigators might examine the potential applications of programs designed to insure individual commitment to organizational goals (i.e., "zero defects," group incentive plans, etc.) to the problem of the commons.

REFERENCES

Bennett, E. B. Discussion, decision, commitment and consensus in "group decision." *Human Relations,* 1955, **8,** 251–274.

Crowe, B. L. The tragedy of the commons revisited. *Science,* 1969, **166,** 1103–1107.

Deutsch, M., & Krauss, M. Studies of interpersonal bargaining. *Journal of Conflict Resolution,* 1962, **6,** 52–67.

Hardin, G. The tragedy of the commons. *Science,* 1968, **162,** 1243–1248.

Kahan, J. P. Noninteraction in an anonymous three-person prisoner's dilemma game. *Behavioral Science,* 1973, **18,** 124–127.

Kelley, H. H., & Grzelak, J. Conflict between individual and common interest in an N-person relationship. *Journal of Personality and Social Psychology,* 1972, **21,** 190–197.

Kelley, H. H., & Stahelski, A. J. The social interaction basis of cooperators' and competitors' beliefs about others. *Journal of Personality and Social Psychology,* 1970, **16,** 66–91.

Likert, R. *The human organization.* New York: McGraw-Hill, 1967.

Platt, J. Social Traps. *American Psychologist,* 1973, **28,** 641–651.

Schachter, S., & Hall, R. Group derived restraints and audience persuasion. *Human Relations,* 1952, **5,** 397–406.

Vinacke, W. E., Mogy, R., Powers, W., Langan, C., & Beck, R. Accommodative strategy and communications in a three-person matrix game. *Journal of Personality and Social Psychology,* 1974, **29,** 509–525.

(Received January 28, 1974)

Journal of Applied Psychology
1974, Vol. 59, No. 6, 717–721

EFFECTS OF THE SHORTER WORKWEEK ON SELECTED SATISFACTION AND PERFORMANCE MEASURES

JOHN M. IVANCEVICH[1]

Department of Business Administration, University of Kentucky

Effects of the 4-day, 40-hour workweek were examined by comparing experimental ($n = 104$) and control ($n = 106$) subjects on measures of job satisfaction, anxiety–stress, and performance over a 13-month period. The analysis indicated that the workers in the 4-day, 40-hour division were more satisfied with personal worth, social affiliation, job security, and pay; experienced less anxiety–stress; and performed better with regard to productivity than their control group (5-day, 40-hour) counterparts.

One of the more recent attempts to improve the job satisfaction and performance of employees is the 4-day, 40-hour (4–40) workweek (see Glickman & Brown, 1973). Basically, the shortened workweek involves more hours of work per day for fewer days per week. It is estimated that between 60 and 70 companies per month are converting to the 4-day workweek (Wheeler, Gurman, & Tarnowieski, 1972). The current estimate of the number of firms using the shorter workweek ranges from 700 to 1,000 (Chard & Caruth, 1973).

The literature suggests that firms are adopting the 4–40 workweek because of potential improvements in production and morale, reduction in absenteeism rates (Steward & Larsen, 1972), and the convenience that it provides employees by providing an extra day off (Poor, 1970). There are, however, some inherent difficulties in the shortened workweek; for example, limitations on the hours of female labor, possible overburdening of supervisory personnel, overtime pay questions, and scheduling problems. More research is needed on the effects of the 4–40 workweek and other workweek variations to clarify the possible benefits and costs of this change in working conditions.

A recent longitudinal study by Nord and Costigan (1973) investigated the reactions of employees over time. There is also a need for researchers to conduct experiments to study the shortened workweek as an independent variable that could influence such dependent variables as satisfaction, anxiety–stress, absenteeism, and performance. The present article describes such an experiment.

The specific hypotheses tested were that when 4-day, 40-hour (4–40) and 5-day, 40-hour (5–40) work groups are compared, the 4–40 subjects show significant improvements in (*a*) self-actualization, (*b*) autonomy, (*c*) personal worth, (*d*) social affiliation, (*e*) job security, and (*f*) pay satisfaction (Hypothesis 1); significant reduction in anxiety–stress (Hypothesis 2); significantly less unexcused absenteeism (Hypothesis 3); significant improvements in performance ratings (Hypothesis 4); and significant changes in the relationships of the job satisfaction facets, anxiety–stress, and absenteeism with overall performance (Hypothesis 5).

METHOD

Company Studied

The company studied has a total of 1,140 managerial and operating employees. These employees are located in manufacturing facilities in four geographically dispersed operations in the Midwest. The four divisions were each engaged in work that could be classified, from a technological standpoint, as intermediately complex (Woodward, 1965). The company designs, manufactures, and markets a packaging machine used in the food-processing industry.

Subjects

The company's executive committee, which wanted to convert all four divisions to the 4–40 workweek, agreed to establish two divisions as experimental and control[2] groups to be studied for 12 months. The experimental division ($n = 104$) was converted to the 4-40 system, while the control group ($n = 106$) continued to function on the 5-40 schedule.

---

[1] Requests for reprints should be sent to John M. Ivancevich who is now at the Department of Organizational Behavior and Management, University of Houston, Houston, Texas 77004.

[2] The term *control division* is used only to indicate the division that was not converted to the 4–40 workweek. The company rejected the request of the author to study the two divisions for at least two years so as to ascertain any "sleeper" effects in the 4–40 division (Campbell & Stanley, 1963).

The experimental and control subjects were similar on such characteristics as mean age, mean seniority, mean time in present grade, unexcused absenteeism rate, education level, and average annual wage. The experimental and control divisions are both located in suburbs of large cities, but are 120 miles apart.

The subjects participating in the study were all members of the United Steelworkers of America Union. Neither division has had a strike for the past 10 years.

Measures

A questionnaire was administered during working hours to the total operating employee population in both divisions. The anonymous questionnaire was completed at three different time periods over a 13-month period. These three time periods are designated (*a*) Time Before, one month before conversion to the 4-40 workweek; (*b*) Time After 1, three months after conversion; and (*c*) Time After 2, 12 months after the conversion.

Job satisfaction was defined as the extent to which the job rewards meet or exceed the employees' perception of an equitable level of rewards on six job facets—self-actualization, autonomy, personal worth, social affiliation, job security, and pay (Porter & Lawler, 1968).

The job satisfaction facet measures were generated from a 19-item questionnaire, previously pretested for clarity and reliability in a third division of the company. Examples of items are the following: (*a*) "There is the opportunity to develop friendships in the company." (*b*) "The compensation program for employees in jobs similar to mine is fair." The respondents answered each of the 19 items on a 1 to 6-point Likert scale (strongly disagree to strongly agree). Acceptable coefficient alpha reliability levels ranging from .70 to .85 were generated from the 19-item job satisfaction measures used.

The *anxiety–stress* measure consisted of nine questions in which employees were asked how frequently they "felt bothered" about a number of job-related matters (see Kahn, Wolfe, Quinn, Snoek, & Rosenthal, 1964). Examples of questions are the following: (*a*) not knowing just what the people you work with expect of you; and (*b*) feeling that you have too heavy a workload, one that you can't possibly finish during an ordinary workday. Employees responded to each item on a 5-point scale (never to nearly all the time).

The *unexcused absence rate* was calculated on the basis of hours per month missed on the job, where the reason was not known by the supervisor.

Four *performance facets* examined were quality, level, effectiveness as a team member, and knowledge of job. The performance index was a multitrait scale, developed by the company, which was completed by the supervisor of each employee. The index was the average of 10 (6-point) ratings evaluating the employee's performance on the four facets relative to "other employees who are now doing the same work" (Spearman–Brown split-half reliability = .81).

Change t tests were used to study the net differences between the changes in the mean scores for both the experimental and the control groups (McNemar, 1962). Although changes in Time After 1 and Time Before, and Time After 2 and Time After 1 scores were analyzed, only the Time After 2 – Time Before net difference changes, the most relevant, are reported in this study.

RESULTS

Table 1 presents data for the evaluation of changes in job satisfaction and anxiety–stress. There was a small, but significant improve-

TABLE 1

MEAN SCORES, NET CHANGES, AND t VALUES FOR SATISFACTION FACETS AND ANXIETY STRESS VARIABLES

Variables	Experimental division ($n = 104$)				Control division ($n = 106$)				Net difference in changes[d]	t value
	TB[a]	TA_1[b]	TA_2[c]	$TA_2 - TB$	TB	TA_1	TA_2	$TA_2 - TB$		
Satisfaction facets										
Self-actualization	3.51	3.58	3.57	.06	3.47	3.46	3.53	.06	.00	0
Autonomy	4.47	4.39	4.48	.01	4.34	4.39	4.38	.04	−.03	.38
Personal worth	4.41	4.49	5.48	1.07	4.42	4.61	4.62	.20	.87	6.79*
Social affiliation	4.18	4.30	4.97	.79	4.40	4.48	4.41	.01	.78	7.01*
Job security	3.61	3.95	4.38	.77	3.72	3.68	3.79	.09	.70	6.99*
Pay	3.04	3.46	4.21	1.17	3.01	3.09	3.10	.09	1.08	10.59*
Overall satisfaction	3.83	4.04	4.75	.92	3.76	3.79	3.81	.05	.87	10.02*
Anxiety–stress	2.61	2.42	1.86	−.75	2.30	2.24	2.25	−.05	−.70	9.11*

Note. The higher the mean score for the first seven variables indicates a larger amount. TB = Time Before; TA_1 = Time After 1; TA_2 = Time After 2.

[a] Mean score of data collected one month before conversion to 4-day, 40-hour workweek.

[b] Mean score of data collected three months after conversion.

[c] Mean score of data collected 12 months after conversion.

[d] The differences in changes for $TA_2 - TB$ period in the experimental and control divisions.

* $p < .01$.

TABLE 2

MEAN SCORES, NET CHANGES, AND *t* VALUES FOR UNEXCUSED ABSENCE AND PERFORMANCE FACTORS

Variables	Experimental division ($n = 104$)				Control division ($n = 106$)				Net difference in changes[d]	*t* values
	TB[a]	TA_1[b]	TA_2[c]	$TA_2 - TB$	TB	TA_1	TA_2	$TA_2 - TB$		
Unexcused Absences	337.0	321.3	331.4	−5.6	349.0	357.2	341.3	−7.7	2.1	.27
Performance facets										
Quality of production	3.83	3.71	4.01	.18	3.75	3.79	3.81	.06	.12	.19
Level of production	3.81	3.99	4.24	.43	3.84	3.81	3.96	.12	.31	3.03*
Effectiveness as team member	3.71	3.94	4.54	.83	3.41	3.62	3.59	.18	.65	9.27*
Knowledge of job	3.31	3.39	3.36	.05	3.93	4.04	3.99	.06	−.01	.09
Overall performance	3.67	3.96	4.04	.37	3.73	3.81	3.84	.11	.26	4.72*

Note. The higher the mean for all of the variables, except unexcused absences, suggests better performance. TB = Time Before; TA = Time After; TA_1 = Time After 1; TA_2 = Time After 2.

[a] Mean score of data collected one month before conversion to 4-day, 40-hour workweek.

[b] Mean score of data collected three months after conversion.

[c] Mean score of data collected 12 months after conversion.

[d] The differences in changes for $TA_2 - TB$ period in the experimental and control divisions.

* $p < .01$.

ment in overall satisfaction in the experimental division which had converted to the 4–40 work schedule. Four of the job satisfaction facets (personal worth, social affiliation, job security, and pay) appeared to account for this general increase in satisfaction. There was no discernible increase in the self-actualization and autonomy facets. In spite of this latter finding, Hypothesis 1 was generally supported.

As predicted in Hypothesis 2, there was a significant (but small) improvement in the perceived anxiety–stress index in the experimental division.

Table 2 presents data relevant to the unexcused absences and performance facets hypotheses. With respect to absences, the results showed no significant changes in either the 4–40 or the 5–40 divisions. Thus, Hypothesis 3 was not supported. The supervisory performance ratings of the employees showed significant improvement for the experimental group in the rate of production, effectiveness as a team member, and overall performance,

TABLE 3

CHANGES IN CORRELATION COEFFICIENTS OF BEHAVIORAL AND ABSENCE VARIABLES WITH PERFORMANCE BETWEEN TIME AFTER 2 AND TIME BEFORE DATA

Variables	Overall performance versus experimental division			Overall performance versus control division			
	TB[a]	TA_2[b]	Difference $TA_2 - TB$	TB	TA_2	Difference $TA_2 - TB$	Net change in correlation[c]
Self-actualization	.17	.09	−.08	.15	.14	−.01	−.07
Autonomy	.13	.07	−.06	.11	.13	.02	−.08
Personal worth	.19	.09	−.10	.13	.04	−.09	−.01
Social affiliation	.18	.46	.28	.12	.19	−.07	.21*
Job security	.01	.54	.53	.09	.01	−.08	.61*
Pay	−.02	.48	.50	−.12	−.07	.05	.45*
Anxiety–stress	−.11	−.19	−.08	−.07	−.02	.05	−.13
Unexcused absences	−.06	−.09	−.03	−.03	−.12	−.09	−.06

Note. Tests of significance were computed after *r* to *z* transformation of correlation coefficients.

[a] Mean score of data collected one month before conversion to 4-day, 40-hour workweek.

[b] Mean score of data collected 12 months after conversion.

[c] The differences in changes for $TA_2 - TB$ period in the experimental and control divisions.

* $p < .01$.

when compared with the control group. Quality and job knowledge ratings showed no differences. Consequently, Hypothesis 4 was only partially supported.

To test Hypothesis 5, which predicted significant changes in the relationships between overall performance and other indices of satisfaction and anxiety, the net correlation changes between the Time After 2 and Time Before were examined for the experimental and control divisions (see Table 3). The results for the experimental groups showed some significantly greater changes than those occurring in the control group. The results suggest that stronger positive relationships between social affiliation, job security, and pay with overall performance occurred in the experimental division after the conversion to the 4–40 workweek. However, there was no significant improvement indicated for the self-actualization, autonomy, personal worth, anxiety–stress, and unexcused absenteeism relationships with overall performance over time. Thus, only partial support of Hypothesis 5 is shown.

DISCUSSION

Overall, the results are generally supportive with regard to positive effects of the 4–40 workweek on a number of the indices investigated. There appeared to be no negative effects on these indices. However, not all indices showed significant effects. The results emphasize the need to evaluate possible effects on a variety of different criteria. Thus, the findings with certain performance facets for the 4–40 division seemed to support the contention of advocates of this schedule that the shorter workweek may result in improved performance (Hall & Nougaim, 1968; Poor, 1970), although not all aspects of performance showed significant change.

Furthermore, the present study indicated that after conversion to the 4–40 workweek, there were positive effects on anxiety levels and on some facets of job satisfaction. However, there was no discernible improvement in satisfaction in autonomy and self-actualization facets. These latter facets are those which Alderfer (1969) labels as "growth" facets. This finding is not consistent with the claims of 4–40 advocates, who assumed positive effects of these facets (Wheeler, Gurman, & Tarnowieski, 1972). Similarly, the lack of a reduction in absenteeism for the 4–40 group does not substantiate previous claims (Nanus & Adelman, 1971; Poor, 1970; Wheeler, 1970).

These results were obtained in two divisions in a single company, and generalizations must be made with caution. Also, the study was limited to a 13-month period, and it is possible that some effects may show up or diminish over a longer time. Nord & Costigan's (1973) results suggest this possibility, and we cannot rule out partial "Hawthorne effects."

The more positive relationships in the experimental division between performance and social affiliation, job security and pay might be partially the result of such an effect. The employees may have considered themselves more secure because they were the only plant in the organization that converted to a relatively modern 4–40 schedule. There are, of course, possible intervening variables, such as physical fatigue, attitudes toward leisure time, job pace, etc. which could be examined as moderators of these effects. Future studies might well examine these issues.

REFERENCES

Alderfer, C. P. A new theory of human needs. *Organizational Behavior and Human Performance*, 1969, 4, 142–175.

Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally, 1963.

Chard, C. A., & Caruth, D. L. A look at the four-day workweek. *Business Studies*, 1973, 12, 40–48.

Glickman, A. S., & Brown, Z. B. *Changing schedules of work: Patterns and implications.* Washington, D. C.: American Institutes For Research, 1973.

Hall, D. T., & Nougaim, K. E. An examination of Maslow's need hierarchy in an organizational setting. *Organizational Behavior and Human Performance*, 1968, 3, 12–35.

Kahn, R. L., Wolfe, D. M., Quinn, R. P., Snoek J. D., & Rosenthal, R. A. *Organizational stress* New York: Wiley, 1964.

McNemar, Q. *Psychological statistics.* New York Wiley, 1962.

Nanus, B., & Adelman, H. M. Work and leisure 1980. *Business Horizons*, 1971, 14, 5–10.

Nord, W. R., & Costigan, R. Worker adjustment t the four-day week: A longitudinal study. *Journa of Applied Psychology*, 1973, 58, 60–66.

Poor, R. *4 Days, 40 Hours.* Cambridge, Mass.: Bursk & Poor Publishing, 1970.

Porter, L. W., & Lawler, E. E., III. *Managerial attitudes and performance.* Homewood, Ill.: Irwin, 1968.

Steward, G. W., & Larsen, J. M. A Four-day—Three-day per week application to a continuous production operation. *Management of Personnel Quarterly,* 1971, 10, 13–20.

Wheeler, K. E. Small business eyes the four-day workweek. *Harvard Business Review,* 1970, **48,** 142–147.

Wheeler, K. E., Gurman, R., & Tarnowieski, D. *The four-day week: An AMA research report.* New York: American Management Association, 1972.

Woodward, J. *Industrial organization.* London: Oxford University Press, 1965.

(Received September 24, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 6, 722–726

SELF-ESTEEM, PERFORMANCE, AND SATISFACTION: SOME TESTS OF A THEORY

JEFFREY H. GREENHAUS [1]
Stevens Institute of Technology

IRWIN J. BADIN
Montclair State College

Two studies were conducted to investigate (*a*) the relationship between the three sources of self-esteem outlined in a 1970 study by Korman—chronic, task specific, and socially influenced—and performance on an anagram task, and (*b*) the moderating effects of task-specific self-esteem on the relationship between performance and satisfaction. College undergraduates participated in each study. It was found that task-specific self-esteem predicted performance for the total sample and socially influenced self-esteem predicted performance for high-authoritarian subjects. It was also found, as predicted, that performance tended to predict satisfaction only for high-self-esteem subjects.

A recent theory of organizational behavior (Korman, 1970) has emphasized the role of self-esteem in work outcomes. Within a balance theoretic framework, the general thrust of the approach is that, all other things being equal, high-self-esteem persons are motivated to perform well on a task in order to maintain their self-image of competence. Low-self-esteem persons, on the other hand, are not motivated to perform well since poor performance for them is consistent with their self-image of relative incompetence.

A particularly unique aspect of Korman's approach is his postulation of three sources of self-esteem. Chronic self-esteem is seen as "a relatively persistent personality trait that occurs relatively consistently across various situations [Korman, 1970, p. 32]." Task-specific self-esteem refers to an individual's feelings of competence for a particular task. Task-specific self-esteem is likely to be a function of past experience with the task or similar tasks. Socially influenced self-esteem is a function of others' expectations of one's behavior. When others expect a person to perform well, and communicate this feeling through their words and behavior, it is expected that the person's own feelings of competence will increase.

The three sources of self-esteem, it is theorized, determine a person's level of "self-perceived competence and ability for the task at hand" which, in turn, directly affects the level of task performance (see Figure 1 in Korman, 1970, p. 35).

The conceptualization of the three sources of self-esteem raises several questions. First, the question of the relative importance that each source of self-esteem has on performance has been neglected. Prior research (see Korman, 1970, for a review of much of the research) has only focused on the effects of each source separately. Second, it would seem that the most potentially useful source of self-esteem from an organization's point of view is socially influenced self-esteem. It would, therefore, be particularly helpful to determine whether practices designed to increase socially influenced self-esteem are more successful for certain types of employees than for others. It is possible, for example, that socially influenced self-esteem should have more of an effect on performance for high-authoritarian persons than for low-authoritarian persons. This prediction is based on two factors: (*a*) that a practice designed to increase self-esteem may be conceived of as a social influence attempt; and (*b*) that there is some evidence (Hovland & Janis, 1959) that high authoritarians are more susceptible to influence attempts than low authoritarians.

In addition to the direct effects self-esteem may have on performance, Korman (1970) has also viewed self-esteem as a moderator of the relationship between performance and satisfaction. He has hypothesized that performance should predict satisfaction only for high-self-esteem persons, whereas satisfaction should predict performance only for low-self-esteem persons.

---

[1] Requests for reprints should be sent to Jeffrey H. Greenhaus, Department of Management Science, Stevens Institute of Technology, Hoboken, New Jersey 07030.

The first component of this moderator effect (performance leads to satisfaction only for high-self-esteem persons) is based on the notion that for high-self-esteem persons, the better the performance, the greater the balance with their self-concept of competence, and therefore the more satisfaction with the task. Because success and competence are not central to the self-concept of low-self-esteem persons their task performance should not have an important impact on their satisfaction. The low-self-esteem person may derive his satisfaction from factors other than effective performance (Greenhaus, 1971).

The second component of the moderator effect (satisfaction predicts subsequent performance only for low-self-esteem persons) is based on a presumed differential reaction of high- and low-self-esteem persons to a state of dissatisfaction with a task. Since it is assumed that individuals seek outcomes that are in balance with their self-concept, high-self-esteem persons will be motivated to achieve a high level of performance on a subsequent task regardless of their satisfaction with a prior task. Considering what has been said about how performance affects satisfaction for high-self-esteem persons, we can expect the high-self-esteem person who is satisfied with a task (primarily resulting from good performance) to continue to perform well on a second task and the high-self-esteem person who is dissatisfied with a task (primarily resulting from poor performance) to improve his performance on a second task to a level about equal to that of the satisfied high-self-esteem person.

For low-self-esteem persons, however, dissatisfaction is presumed to result in the stereotypically poor performance of the disgruntled employee (Korman, 1970). Therefore, we would expect a decrease in the performance of dissatisfied low-self-esteem persons and no change in the performance of satisfied low-self-esteem persons, that is, a positive correlation between satisfaction and performance.

Two studies were conducted to investigate the variety of predictions outlined above. The first study was concerned with the main effects of self-esteem. It was predicted (Hypothesis 1) that there would be positive relationships between each of the three sources of self-esteem and task performance. It was also predicted (Hypothesis 2) that the relationship between socially influenced self-esteem and task performance would be greater for high-authoritarian subjects than for low-authoritarian subjects. In the second study it was predicted (Hypothesis 3) that performance would be positively related to satisfaction with the task only for high-self-esteem subjects, whereas (Hypothesis 4) task satisfaction would be positively related to performance on a second task only for low-self-esteem persons.

EXPERIMENT 1

Method

Subjects

The subjects in this study were 144 undergraduate students (75% male) enrolled in four sections of industrial and organizational psychology at an eastern state college.

Research Instruments

Chronic self-esteem was measured by the self-assurance scale of the Self Description Inventory (Ghiselli, 1971). Task-specific self-esteem was measured by five items, all intended to tap the subject's self-esteem for a particular task, solving anagrams. Specifically, the subject rated (*a*) his ability in solving anagrams (compared to solving arithmetic and mechanical problems); (*b*) his *aptitude* for solving anagrams; (*c*) an estimate of what his performance might be on an anagram task; and (*d*) his *capability* in solving anagrams. The (alpha) reliability of the task-specific self-esteem scale was .72.

Socially influenced self-esteem was manipulated through differential instructions given to describe the nature of the study. Specifically, the instructions given to high socially influenced self-esteem subjects were as follows:

> The task you are about to begin has been given to a nation-wide sample of people. It has been found that college students in New Jersey like yourself do very well on this task. In fact, they do well above the national average. I expect that you will also do rather well on this task.

The instructions given to low socially influenced self-esteem subjects were as follows:

> The task you are about to begin has been given to a nation-wide sample of people. It has been found that college students in New Jersey like

> yourself do not do very well on this task. In fact, they do well below the national average. I expect you will also do rather poorly on this task. I am asking you to take it, however, so that we can add more data to the national sample.

The correlations involving socially influenced self-esteem were product-moment correlations with values of one and two representing low and high self-esteem, respectively.

Authoritarianism was measured by Form 40 of the F scale (Adorno, Frenkel-Brunswik, Levinson, & Sanford, 1950). In developing low- and high-authoritarian subgroups, the sample was dichotomized at its median (112). The performance task used consisted of 20 5-letter anagrams taken from Mayzner and Tresselt's (1958) list of difficult sequences. The criterion of task performance was the number of anagrams successfully completed in a 15-minute time period.

Procedure

Two of the participating classes ($n = 67$) were assigned to the high socially influenced self-esteem condition and the two remaining classes ($n = 77$) were assigned to the low socially influenced self-esteem condition. Following the differential instructions (depending on the experimental condition), each subject received an identical questionnaire booklet which contained in the following order: chronic self-esteem scale, task-specific self-esteem scale, F scale, and anagram task. After completing the booklet, subjects were told the true nature of the study.

Results

Table 1 presents the intercorrelations among the major variables of the study. It should be noted, first, that the three sources of self-esteem were virtually uncorrelated. Also, regarding Hypothesis 1, only task-specific self-esteem was significantly ($p < .01$) related to task performance.

TABLE 1

CORRELATIONS BETWEEN SOURCES OF SELF-ESTEEM AND TASK PERFORMANCE ($N = 144$)

Variable	Chronic self-esteem	Socially-influenced self-esteem	Task performance
Task-specific self-esteem	−.03	−.09	.31*
Chronic self-esteem	—	−.10	−.16
Socially-influenced self-esteem		—	.09

* $p < .01$.

The data provide support for Hypothesis 2. The correlation between socially influenced self-esteem and performance for the high authoritarians was .29 ($n = 70$, $p < .02$) and for the low authoritarians was −.02 ($n = 74$, *ns*). The difference between the correlations was signficant ($z = 1.88$, $p < .05$, one-tailed).

Discussion

Task-specific self-esteem was the only significant determinant of performance. However, as predicted, the relationship between socially influenced self-esteem and performance was significant for high-authoritarian subjects. This finding may have several implications. First, an organization might decide to concentrate its self-esteem intervention techniques on high-authoritarian employees where there is a higher probability of success; or, alternatively, different techniques might be considered for low-authoritarian employees. In a theoretical vein, it may be useful to study socially influenced self-esteem as a particular type of social influence attempt. In this way, subsequent attitude (and performance) change could be investigated in light of such variables as the characteristics of the communication and communicator, the personality predisposition of the *audience* and the potential influence of group pressure.

EXPERIMENT 2

Method

Subjects in the second study were 61 undergraduates in three sections of industrial and organizational psychology at an eastern state college. The second study concentrated on the moderator effects of task-specific self-esteem since it was the best predictor of performance in the first study. Task-specific self-esteem was measured by the same scale used in the first study. In order to test the impact of performance on satisfaction and satisfaction on later performance, it was necessary to develop two similar tasks separated by a measure of task satisfaction. The original 20-item anagram task used in the first study was, therefore, separated into two 10-item tasks. The 10 odd-numbered anagrams constituted the first task and the 10 even-numbered anagrams were used as the second task. Task satisfaction was measured by the subject's response to the question, "How much did you like doing the anagram task?" Responses were made on a 5-point scale from "liked it very much" to "disliked it very much."

Subjects received a booklet which contained in the following order: task-specific self-esteem scale, Task

TABLE 2

CHANGES IN PERFORMANCE FROM TASK 1 TO TASK 2 AS A FUNCTION OF SATISFACTION WITH TASK 1 AND TASK-SPECIFIC SELF-ESTEEM

Task-specific self-esteem	Satisfaction with Task 1	n	Task 1 Performance $\bar{X}$	Task 1 Performance SD	Prediction of change in performance	Task 2 Performance $\bar{X}$	Task 2 Performance SD	t
High	Satisfaction	20	4.45	2.24	Remain same	3.00	2.59	2.90*
High	Dissatisfaction	17	2.47	1.41	Increase	2.24	1.89	<1
Low	Satisfaction	11	3.82	1.99	Remain same	2.91	1.92	1.71
Low	Dissatisfaction	13	3.00	1.87	Decrease	1.69	1.44	3.59*
Total		61	3.48	2.04		2.49	2.10	

*$p < .01$.

1, scale of satisfaction with Task 1, and Task 2. After subjects completed the first task (but before they rated their satisfaction), the experimenter read the correct answer to each anagram and, in order to provide feedback on their relative success, told subjects that a score of 6 or more correct meant they had "passed" the test and a score of 5 or less meant that they had "failed" the test. (These cutoffs were based on score distributions on a pretest of the tasks.) After hearing whether they passed or failed the tasks, subjects rated their satisfaction with the task and went on to the second anagram task.

Results

In order to test Hypothesis 3, subjects were dichotomized into high task-specific ($n = 37$) and low task-specific ($n = 24$) self-esteem subgroups. (The dichotomy was approximately a 60%–40% split because the distribution did not permit a split exactly at the median, and a random allocation of subjects near the median into the high and low subgroups was rejected.) Then, within each self-esteem subgroup, correlations between performance (raw scores) on Task 1 and satisfaction with Task 1 were computed. The performance–satisfaction correlation for the high task-specific subjects was .50 ($p < .01$) and for the low task-specific self-esteem subjects was .17 (*ns*). However, the difference between the two correlations ($z = 1.4$) was nonsignificant.

Hypothesis 4 predicted that satisfaction should correlate with subsequent performance for low-self-esteem persons but not for high-self-esteem persons. Within each level of task-specific self-esteem, correlations between satisfaction with Task 1 and performance on Task 2 were computed. For low task-specific self-esteem subjects, the satisfaction–performance correlation was .37 ($p < .05$, one-tailed), whereas for high task-specific self-esteem subjects the correlation was .20 (*ns*). The difference between the correlations was nonsignificant.

Table 2 presents changes in performance from Task 1 to Task 2 for subjects subgrouped on task-specific self-esteem and satisfaction with Task 1. Perhaps the most obvious finding is that all of the subgroups decreased in performance from Task 1 to Task 2. An inspection of the performance changes of the four subgroups provides little support for Hypothesis 4.

It might be noted parenthetically that although Experiment 2 concentrated on the moderator effects of task-specific self-esteem because it significantly predicted performance in Experiment 1, the results for Experiment 2 were not so strong. In fact, the correlation between task-specific self-esteem and performance in the latter experiment was .14 (compared to an r of .31 in Experiment 1). It should also be noted that there was a restriction of range on task-specific self-esteem in Experiment 2 ($SD = 2.42$) compared to Experiment 1 ($SD = 3.14$).

Discussion

The data in Experiment 2 lend tentative support to at least one of the moderating effects suggested by Korman (1970). Performance did influence satisfaction only for high-self-esteem subjects. This result not only tends to confirm Korman's (1968) earlier finding but extends the generality of the hypothesis since in this study task-specific and not chronic self-esteem served as the moderator. It should be remembered, however, that the difference between the performance–satisfaction correlations for high- and low-self-esteem subjects was nonsignificant. The other moderator effect, satisfaction predicts performance only for low-self-esteem subjects, received little support.

It should be noted that the measure of satisfaction used in the present study was specific to the task. Perhaps future research in industrial settings would benefit by using either a more general measure (e.g., Brayfield–Rothe) or a multifactor measure (e.g., Job Descriptive Index) of satisfaction.

OVERALL COMMENTS

The results of the two studies suggest the important role self-esteem may play in shaping task performance and satisfaction. The main effect of task-specific self-esteem in Experiment 1 may have implications for organizational practice. Assuming that the possession of relevant abilities is at least moderately related to task-specific self-esteem, effective selection, classification, and training systems can generate a work force with a relatively high degree of self-perceived competence. Perhaps any experience that provides for skill acquisition, mastery, and psychological growth should contribute to employees' level of task-specific self-esteem.

Task-specific self-esteem seems related to other notions of expectancy (Porter & Lawler, 1968; Vroom, 1964) since they all refer to perceived probabilities of effective performance. Since expectancy theory proposes an interactive relationship between expectancy and valence, to the extent that task-specific self-esteem is raised, improvements in an organization's reward structure (leading to a higher valence of effective performance) should result in larger increments in work motivation.

More research should also be conducted on the interrelationships among the three sources of self-esteem over time. It might be expected, for example, that a *steady dose* of high socially influenced self-esteem would result in higher levels of employees' chronic and task-specific self-esteem. That is, leadership practices of the Theory Y variety could foster both general positive feelings (I am competent, I can handle responsibility) as well as self-perceptions of competence in the particular set of activities in which the employees are engaged. The required intensity and duration of the socially influenced self-esteem attempt are, of course, important empirical issues.

REFERENCES

Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., & Sanford, R. N. *The authoritarian personality.* New York: Harper, 1950.

Ghiselli, E. E. *Explorations in managerial talent.* Pacific Palisades, Calif.: Goodyear Publishing, 1971.

Greenhaus, J. H. Self-esteem as an influence on occupational choice and occupational satisfaction. *Journal of Vocational Behavior,* 1971, 1, 75–83.

Hovland, C. I., & Janis, I. L. (Eds.) *Personality and persuasibility.* New Haven: Yale University Press, 1959.

Korman, A. K. Task success, task popularity, and self-esteem as influences on task liking. *Journal of Applied Psychology,* 1968, 52, 484–490.

Korman, A. K. Toward an hypothesis of work behavior. *Journal of Applied Psychology,* 1970, 54, 31–41.

Mayzner, M. S., & Tresselt, M. E. Anagram solution times: A function of letter order and word frequency. *Journal of Experimental Psychology,* 1958, 56, 376–379.

Porter, L. W., & Lawler, E. E. *Managerial attitudes and performance.* Homewood, Ill.: Dorsey Press, 1968.

Vroom, V. H. *Work and motivation.* New York: Wiley, 1964.

(Received December 11, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 6, 727–730

NOTE ON THE UTILITY OF CERTAINTY WEIGHTS IN EXPECTANCY THEORY[1]

JACK M. FELDMAN[2]

Department of Management, University of Florida

The usefulness of "certainty of response" as a moderator variable in expectancy models was tested in the context of Fishbein's construct of "attitude toward the act." The direct evaluation of five job-seeking behaviors by a racially and socioeconomically heterogeneous sample ($N = 200$) was predicted from five combinations of outcome evaluation, outcome expectancy, and response certainty. Contrary to the hypothesis, the simple sum of the evaluations of the outcomes of each behavior best predicted the directly measured evaluation. Results are discussed in terms of possible modifications of the expectancy model and the concept of uncertainty as reflected in subjective probability estimates.

Instrumentality–expectancy theory, as formulated by Vroom (1964), Porter and Lawler (1968), Graen (1969), Fishbein (1967), and others, has as a basic tenet that behavior is a multiplicative function of the expected consequences of an action and the individual's evaluation of those consequences. While details of the theories differ, an expectancy x valence term is common to all, whether it is called "force" (Vroom, 1964), "attitude toward the act" (Fishbein, 1967), or "internal pressure" (Graen, 1969).

Graen (1969) has, however, found that certain boundary conditions limit the generality of expectancy theory. Specifically, he found that behavior was predictable from expectancy theory only in a "reciprocating climate"—where performance is regularly evaluated, good performance rewarded, and the relevant contingencies communicated to the worker. Camman and Lawler (1973) reported data consistent with this finding, though the proposition was not directly tested.

A reciprocating climate is taken to be a climate in which stable subjective probabilities (expectancies) can be formed. If reinforcement contingencies continually change in a fashion perceived by the workers to be entirely random—for example, where a department is continually being reorganized—then stable estimates of behavior–reward probabilities will most likely not be formed. According to Graen (G. Graen, personal communication, 1971) expectancy models cannot predict attitudes or behavior if people cannot form stable expectancies, as in a nonreciprocating climate. This perceived instability ought to be reflected in the certainty statements given for the expectancy measures. Thus, Graen suggests the use of a "certainty of response" measure, used to weight the expectancy estimate, in the computation of internal pressure.

The use of certainty weights ought to increase the validity of the model, since more certain responses would receive more weight in the equation. This suggestion is similar to Brigham's (1971) proposal that the validity of stereotype measures in the prediction of interpersonal behavior could be improved by the use of a certainty weight. Some support is provided by Sample and Warland (1973), who found that student-election voting behavior could be more accurately predicted from attitudinal measures for students who were highly certain of their responses.

Fishbein's (1967) construct of attitude toward the act can be used as the basis for a test of the certainty hypothesis. Fishbein

---

[1] This article is based on part of a dissertation submitted in partial fulfillment of the requirements for the doctoral degree at the University of Illinois. The research was supported by the Social and Rehabilitation Service of the U.S. Department of Health, Education, and Welfare (Research Grant 15-P-55175/5; Harry C. Triandis, principal investigator). Thanks are due to Joe Reitz and Wally Hill for their comments on an earlier version of this manuscript and to Bob Hilterman and Jean Feingold for their help in data analysis.

[2] Requests for reprints should be addressed to Jack M. Feldman, Department of Management, University of Florida, Gainesville, Florida 32611.

(1967) defines attitude toward the act as the evaluation of an act in a situation; theoretically, it combines in a linear multiple regression equation with a measure of normative beliefs about a behavior, to predict the intention to perform a particular act. These "behavioral intentions" mediate between the model and measures of overt behavior. (For further discussion and a review of the evidence pertaining to the model, see Ajzen and Fishbein, 1973.)

Attitude toward the act is a multiplicative function of beliefs about the consequences of an act (B_i) and the evaluation of those consequences (a_i). In the model, B_i corresponds to an expectancy, or subjective probability, that a behavior will be followed by an outcome. As Feldman (1973) has pointed out, attitude toward the act and Graen's (1969) concept of internal pressure seem to be virtually identical.

Given the isomorphism of attitude toward the act and internal pressure, it is likely that anything affecting the validity of an expectancy estimate would also affect the relationship between the theoretical estimate of attitude toward the act, $\sum_{i=1}^{N} (B_ia_i)$, and a direct measure of the evaluation of an action. Thus, if a certainty of response measure does reflect the degree to which a subject places confidence in his probability estimates, then a multiplicative certainty weight for each B_i should improve the correlation between the above theoretical estimate and direct measure of attitude toward the act.[3] Denoting certainty as C_i, the question becomes: Which correlates more highly with a direct measure of attitude toward the act, $\Sigma(B_ia_i)$ or $\Sigma(C_iB_ia_i)$?

However, this question ignores the issue of whether sum of products models generally provide the best estimates of directly measured evaluation. To assess the validity of the expectancy model in general, it is necessary to compare the correlations obtained by several different combinations of the following variables: expectancy (B_i), outcome–evaluation (a_i), and certainty (C_i). It is hypothesized, therefore, that $\Sigma(C_iB_ia_i)$ will correlate most highly with directly measured attitude toward the act, followed by $\Sigma(B_ia_i)$, Σa_i, $\Sigma(C_iB_i)$, and ΣB_i. This prediction is based on the assumption that the above combinations contain successively decreasing amounts of information.

METHOD[4]

Subjects

Subjects were 200 male, paid volunteers, between the ages of 18 and 50. Approximately 50 subjects fell into each of the following categories: black hard-core unemployed, black working class, white hard-core unemployed, and white working class.

Measures and Procedure

The evaluations of five job-seeking behaviors were used in testing the hypothesis. They were (*a*) going to an employment agency (agency), (*b*) going from place to place filling out applications (applications), (*c*) trying to join a training program (training), (*d*) reading want ads (ads), and (*e*) hanging around with friends until something turns up (friends). These behaviors had been most frequently mentioned by a pretest sample of blacks and whites as typical job-seeking activities.

Following Graen (1969), intrinsic outcomes of each behavior were also elicited from the pretest sample. "Intrinsic" in this case means intrinsic *to the behavior itself*. Subjects were asked to "name things that happen to you just because you (for example) go to an employment agency." Thus, "getting a job" would *not* be an intrinsic outcome of going to an employment agency, but "filling out long applications," "waiting for an appointment," etc., might be. The five intrinsic outcomes of each behavior most frequently mentioned by the pretest sample were used.

Evaluations of each of the 25 outcomes and five behaviors were made on the nine-alternative form of Kunin's (1955) Faces Scale, presented on a flashcard. This scale provides an approximately equal-interval measure tapping an affective response dimension (Feldman, 1972).

Expectancy (subjective probability) was measured by asking each subject to estimate the chances in 10 of receiving each outcome, having done the behavior. Subjects responded as if they were setting odds on a bet. No visual rating scale was used.

---

[3] Note that, to the extent that some theoretical estimate of the attitude toward the act correlates with a direct measure of the attitude toward the act for a given behavior, the convergent validity of attitude toward the act as a hypothetical construct has been demonstrated.

[4] This methodology section is necessarily brief. Interested readers are referred to Feldman (1973) where full details of the procedure, subject population, etc. are presented.

Certainty was measured after each expectancy response. Subjects were shown a flashcard with a three-alternative scale, ranging from "very sure this is right" ( ) to "don't know at all" (1), and chose the response that best represented their confidence in the expectancy estimate.

Subjects were familiarized with the task and given practice items before data collection began. Those who could not comprehend the task after practice were excused. Data were collected as part of a longer interview, two to four hours long. Interviewers were of the same race and sex as the subjects and had been trained in neutral tones of voice and mannerisms. Full details of the interview procedure, practice items, scales, interview schedules, and validity data on the measures can be found in Feldman (1972).

RESULTS AND DISCUSSION

Since race and social class differences are irrelevant to the hypothesis, analyses were performed on the combined sample ($N = 200$). The directly measured evaluation of each behavior (attitude toward the act) was correlated with each of five predictors:

1. the sum of the expectancies of obtaining each of the five outcomes of each behavior (ΣB_i);
2. the sum of the products of each expectancy times its corresponding certainty weight, taken across the five intrinsic outcomes of each behavior $[\Sigma(C_iB_i)]$;
3. the sum of the evaluations of the five intrinsic outcomes of each behavior (Σa_i);
4. the sum of the products of evaluation of each outcome times the expectancy of obtaining that outcome $[\Sigma(B_ia_i)]$ across the five outcomes of each behavior;
5. the sum of the products of expectancy times certainty times evaluation of each outcome, taken across the five outcomes of each behavior $[\Sigma(C_iB_ia_i)]$.

Thus, the hypothesis was tested five times, once for each of the five job-seeking behaviors. These tests are not independent, however, since the same subjects responded to all items.

As Table 1 shows, the hypothesis was not supported. The best predictor of directly measured attitude toward the act was the simple sum of the evaluations of the intrinsic outcomes of each behavior, $\Sigma(a_i)$. The addition of expectancy as a multiplicative weight, $\Sigma(B_ia_i)$, reduced the obtained correlation in every case, significantly so for all behaviors except training (agency, $p < .05$; applications, $p < .01$; ads, $p < .01$; friends, $p < .05$). The addition of multiplicative certainty weights, $\Sigma(C_iB_ia_i)$, further reduced the correlations, though the differences between $\Sigma(B_ia_i)$ and $\Sigma(C_iB_ia_i)$ as predictors of attitude toward the act were not significant for any behavior.[5]

TABLE 1

CORRELATIONS OF ALTERNATIVE PREDICTION MODELS WITH DIRECTLY MEASURED EVALUATION OF FIVE JOB-SEEKING BEHAVIORS

Predictor	Behavior				
	Agency	Applications	Training	Ads	Friends
ΣB_i	.18*	−.20**	−.24***	−.13	−.10
$\Sigma(C_iB_i)$	.14*	−.18*	−.31***	−.25***	−.10
Σa_i	.75***	.60***	.58***	.44***	.29***
$\Sigma(B_ia_i)$	.68***	.45***	.53***	.27***	.15*
$\Sigma(C_iB_ia_i)$	.64***	.42***	.48***	.12	.14*

Note. $N = 200$. All correlations are rounded to the second decimal. Negative correlations reflect the negative evaluations of several outcomes for each behavior. These signs are naturally reversed when the evaluations are specifically accounted for in the last three prediction models.

* $p < .05$.
** $p < .01$.
*** $p < .001$.

Though attitude toward the act appears to be a meaningful construct, predictable from the sum of the evaluations of the act's intrinsic outcomes, the underlying expectancy model may need modifications. Kanouse and Hanson (1971), for example, cite a number of studies in the risk-taking area which support the proposition that the reinforcement value of a potential loss is a more important determinant of behavior than the probability of the loss itself. Since several of the outcomes of each behavior in the current study were evaluatively negative, the same process may be operative in decision making under uncertainty. That is, even where an actual risk is not apparent, people may act to avoid a potential loss (in this case, of time or opportunity). This avoidance would be reflected in their evaluation of the behavior, as in the data of Table 1.

---

[5] These results replicate across different attitude scales, though some method variance is present. A discussion of race and socioeconomic class effects on these relationships is found in Feldman (1972, 1973).

It appears from these data that certainty weighting decreases the validity of the expectancy model. Two possible explanations should be mentioned.

First, certainty may already be represented in the expectancy estimate. Some intermediate expectancy may represent maximal uncertainty, while deviations from this point in either direction represent increasing certainty. This would suggest the use of a bipolar measure such as Fishbein's (1967), rather than a direct probability estimate, in future research.

Second, certainty may not be relevant to the measurement of subjective probabilities, since the individual is already assumed to be giving his "best guess" as to the state of the world. Certainty might be more relevant to statements such as "Current economic policy is beneficial" than to statements of the form "The chances are 1 in 10 that inflation will decrease this month."

It appears from a closer analysis of the data that both of the above explanations may be valid, the former somewhat less frequently than the latter. Polynomial regression analyses were run on each of the 25 outcomes of the five behaviors, testing the significance of the linear, quadradic, and cubic components of variance in the regression of expectancy on certainty. In each of 25 cases, the linear component was significant at or beyond the .01 level. In 10 of 25 cases, the quadradic component was significant at or beyond the .05 level, generally controlling less variance than the linear component. The cubic component was significant ($p < .05$) in only 3 of the 25 cases. Generally, where the quadradic component was significant, the lowest degree of certainty occurred between 2 and 4 on the expectancy scale, although in two cases the inflection point was close to 7.[6]

Two general conclusions may be drawn. First, certainty weights for subjective probability estimates are not likely to prove useful, either in terms of predictive validity or in aiding our understanding of attitude formation and decision making. Where certainty weights have proved useful, they have been operationally similar to subjective probability judgments (i.e., Sample & Warland, 1973). Second, subjective probabilities (expectancies) may well be irrelevant to the formation of attitudes toward behaviors (and, by implication, behavioral intentions and actual decisions) where negative outcomes of a behavior exist. Further studies in both laboratory and field settings are needed to uncover the complex interactions suggested by the present data.

REFERENCES

Ajzen, I., & Fishbein, M. Attitudinal and normative variables as predictors of specific behaviors. *Journal of Personality and Social Psychology*, 1973, **27**, 41–57.

Brigham, J. C. Ethnic stereotypes. *Psychological Bulletin*, 1971, **76**, 15–38.

Camman, C., & Lawler, E. E. Employee reactions to a pay incentive plan. *Journal of Applied Psychology*, 1973, **58**, 163–172.

Feldman, J. M. Race, economic class, and job-seeking behavior: An exploratory study. Unpublished doctoral dissertation, University of Illinois, 1972.

Feldman, J. M. Race, economic class, and job-seeking behavior: An exploratory study. *JSAS Catalog of Selected Documents in Psychology*, 1973, **3**, 99. (Ms. No. 438)

Fishbein, M. Attitude and the prediction of behavior. In M. Fishbein (Ed.), *Attitude theory and measurement*. New York: Wiley, 1967.

Graen, G. Instrumentality theory of work motivation: Some experimental results and suggested modifications. *Journal of Applied Psychology Monograph*, 1969, **53**(2, Pt. 2)

Kanouse, D. E. & Hanson, L. R. Negativity in evaluations. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, and B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior*. Morristown, N. J.: General Learning Press, 1971.

Kunin, T. The construction of a new type of attitude measure. *Personnel Psychology*, 1955, **8**, 65–77.

Porter, L. W., & Lawler, E. E. *Managerial attitudes and performance*. Homewood, Ill.: Irwin, 1968.

Sample, J., & Warland, R. Attitude and prediction of behavior. *Social Forces*, 1973, **51**, 292–304.

Vroom, V. *Work and motivation*. New York: Wiley, 1964.

(Received December 17, 1973)

---

[6] Complete data, omitted for reasons of space, may be obtained from the author.

Journal of Applied Psychology
1974, Vol. 59, No. 6, 731–740

AN ANALYSIS OF SITUATIONAL MODERATING EFFECTS ON THE RELATIONSHIPS BETWEEN LEAST PREFERRED CO-WORKER AND OTHER PSYCHOLOGICAL MEASURES[1]

MARSHALL SASHKIN[2]

Department of Management and Organization Sciences, School of Business Administration, Wayne State University

F. CARTER TAYLOR[3] AND RAMA C. TRIPATHI[4]

Survey Research Center, Institute for Social Research, Univeristy of Michigan

Correlations were computed between a manager's Least Preferred Co-worker (LPC) score and 17 psychological measures and moderated by subordinates' satisfaction with the manager in 34 offices of a nationwide brokerage firm. The results tended to support the hypothesis that correlations between LPC and the psychological measures should be opposite in direction for high- versus low-satisfaction offices. The hypothesis was directionally confirmed for 11 of 17 pairs of correlations, and the differences were statistically significant for 5 of the 17 pairs. The results also support Fiedler's interpretations of effective leader behavior and the relation between LPC and behavior.

Fiedler's (1967) contingency model of leadership effectiveness is one of the most extensively researched of all current approaches to leadership. Much experimental research in laboratory and field settings (Chemers & Skrzypek, 1972; Fiedler, 1968, 1971a, 1971c; Fiedler & Chemers, 1974) has shown that in certain types of situations, the two types of leader—task motivated, or *low* LPC (least preferred co-worker), and relationship motivated, or ***high*** LPC—are differentially effective.

Several studies (Graham, 1968, 1973; Sample & Wilson, 1965; Sashkin, 1971, 1972), when examined together, suggest possible explanatory mechanisms for the differential effectiveness of high- and low-LPC leaders in the different situations.

Fiedler (1967) created a set of eight situation types, ranging along a postulated scale of "favorableness for the exercise of leadership," that is, how easy or difficult it would seem for a leader to get subordinates to do what the leader wants. When the situation is high in favorableness (Types 1, 2, and 3), the low-LPC leader performs better than the high-LPC leader. What seems to happen is that the low-LPC leader perceives the "ease" of the situation and the liklihood of task attainment. Thus, feeling secure that his needs will be met, the leader feels able to attend to interpersonal relationship factors. This is done by using interpersonal, "socio-emotional" behaviors to reinforce subordinate task behaviors which seem appropriate, while maintaining a normal level of task structuring or directive behavior. The outcome is effective group performance. The high-LPC leader uses a similar implicit logic. To this leader the situation "looks good," in general, in terms of meeting the leader's needs, since leader–member relations in very favorable situations are quite good. The high-LPC leader can, therefore, concentrate on task factors.

This is done by using directive, "initiating," or "structuring" behaviors, while maintaining a normal, general level of relationship-oriented behaviors. The result is poor performance, since (*a*) the structuring behaviors are not rewarding for subordinates, and (*b*) the appropriate focus for the leader is on relationship-oriented behavior, since the task is, by the nature of the situation, likely to be progressing smoothly.

---

[1] This research was supported by the firm in which the research was conducted while the latter two authors were members of the Institute for Social Research, University of Michigan.

We are grateful for the cooperation of the members of the organization studied and for the helpful advice of Stanley E. Seashore.

[2] Requests for reprints should be addressed to Marshall Sashkin, Department of Management and Organization Sciences, School of Business Administration, Wayne State University, Detroit, Michigan 48202.

[3] Now at the Graduate School of Business Administration, Kent State University.

[4] Now at the University of Allahabad, Allahabad, India.

In situations of moderate favorableness (Types 4 and 5) or moderate unfavorableness (Types 6 and 7), high-LPC leaders are found to be superior to low-LPC leaders. According to Fiedler (1971a), the leader in such moderate situations becomes aware that his needs may not be satisfied. The high-LPC leader responds by focusing behaviors on interpersonal relationships, the primary means this type of leader has developed for attaining need satisfaction. Since a major factor in these moderate situations is, supposedly, problems concerning interpersonal relations (Fiedler, 1971a, p. 17), the result is good performance. The low-LPC leader becomes similarly concerned, and responds analogously, by focusing on task-structuring or directive behaviors. This focus, however, is not appropriate, and the outcome is poor performance.

Finally, there are situations so grossly unfavorable (Type 8) that only by highly directive leadership behavior can anything at all be accomplished. The low-LPC leader, engaging in directive, perhaps authoritarian, behaviors then proves more effective than high-LPC leaders, although the results may be only temporary.

Fiedler's theory and empirical model have not been without criticism, on a variety of grounds, primarily (*a*) the theoretical logic (McMahon, 1972), (*b*) the methodology of his research (Graen, Alvares, Orris, & Martella, 1970), (*c*) the validity of his results (Graen, Orris, & Alvares, 1971), and (*d*) the utility of the model (Duffy, Kavanagh, MacKinney, Wolins, & Lyons, 1970; Sashkin, 1972). While these issues are to a degree interrelated, it is useful to discuss each one briefly.

Theory

McMahon (1972) has presented an excellent and thorough review of theoretical problems. One of the more serious of these concerns the nature and meaning of the LPC measure. Fiedler's interpretation is based on logic, not evidence. While he initially avoided any claim that LPC measures some personality attribute, Fiedler (1972; Fiedler & Chemers, 1974) has more recently asserted that LPC yields "an index of the hierarchy of goals and needs" of the individual, an important personality variable "which is not measured by the usual psychological tests [Fiedler & Chemers, 1974, p. 99]."

The problem is that LPC is related to no other psychological measures, usual or unusual. Fiedler and others have been consistently unable to show any meaningful relationships between LPC scores and any of a wide variety of commonly used psychological measures (Fiedler, 1967, pp. 278–280) which might logically relate to an individual's motivational hierarchy or task versus relationship orientation. While Mitchell (1970) did demonstrate some consistent relationships between LPC and a measure of cognitive complexity very similar to that developed by Scott (1962), it is not clear that cognitive complexity is, in itself, a particularly stable factor. Furthermore, the construct and concept of cognitive complexity is not unrelated to the LPC measure. Going back to Fiedler's (1967) earliest work in this area, which centered on the leader's ability to differentiate among subordinates, one finds logical as well as concrete developmental ties between LPC and cognitive complexity. Even so, a recent study by Larson and Rowland (1974) failed to replicate Mitchell's findings.

One other attempt to interpret the LPC measure is found in the report by Fishbein, Landy, and Hatch (1969). Using an attitudinal analysis, these authors seemed to suggest that LPC may measure an individual's attitude toward some specific other individual—a least preferred co-worker—and nothing more and provided some support for this position. In sum, Fiedler's (1967) note that "we were in the peculiar position of having a score which seemed to correlate with nothing but group performance [p. 46]" holds almost as well today as when written.

Research Methodology

Serious methodological issues have been raised by certain critics of Fiedler's work (e.g., Graen et al., 1970, 1971). Fiedler (1971b, 1971c), however, presented detailed and thorough counterarguments. These presentations, as well as tightly designed and controlled laboratory studies (e.g., Sashkin, 1971, 1972) which meet Fiedler's (1971b) research specifications and which yielded strong support for the contingency model, lead us to conclude that Fiedler's research methods—the basic design and approach, not necessarily specific measures or studies—adequately meet the requirements for rigorous, scientific research.

Validity

The third problem area deals with the empirical validity of the contingency model. Graen et al. (1970, 1971) labeled this an issue of "antecedent versus evidential" validity. That is, the model was derived empirically from the numerous experimental results, but does it, in fact, have predictive validity? While Graen et al. (1971) attempted to show that it does not, Fiedler (1971b) has pointed out serious flaws in the research design and method which Graen et al. used in their evidential tests. The predictive test evidence presented by Fiedler (1971c) and others (e.g., Chemers & Skrzypek, 1972; Fiedler & Chemers, 1974; Sashkin, 1972) in field and laboratory studies indicates quite clearly that the model does in fact predict leadership effectiveness, although to varying degrees, depending on which of the eight situations one refers to.

Utility

Finally, there is clearly a question of utility, or practicality. Even if we assume that the model is fully accurate and predicts precisely as it states, the median correlations between LPC and group performance, for each of the eight situations, range between +.13 and +.59 (Fiedler & Chemers, 1974). In terms of the variance in performance attributable to LPC score, the percentages range from about 0%, at worst, to about 36% at best. If we assume that actual situations fall in equal proportion into the eight types (which is, obviously, ridiculous), the model would, on the average, account for about 11% of the variance in group performance. While this is nothing to scorn, neither is it a radical breakthrough in social engineering, particularly since Duffy et al. (1970) showed in a field study that the predictive validity of the model is seriously reduced when performance is measured in terms of supervisory ratings rather than "hard" performance criteria.

THE PRESENT STUDY

Issues and Aims

We are concerned here primarily with the first of the problems reviewed above and, to a lesser extent, with the third and fourth. We shall briefly detail our approach to these issues.

What does LPC measure? Given the evidence or, more appropriately, the lack thereof, Fiedler's interpretation of the LPC measure must be considered speculative, at best. Yet, LPC must measure—and mean—something, for it seems reasonably clear that leaders' LPC scores are predictive of group performance to a consistent, and in many cases significant, degree. Examination of relationships between the measure of interest and a variety of other measures is one common approach to understanding the meaning of a psychological measure. This approach, however, has yielded little in the way of explanation or understanding of the LPC measure.

We earlier summarized the results of several studies which suggest explanatory mechanisms for Fiedler's model and discussed in detail some of this research when reviewing the problems associated with the model. Two reports in particular serve as the background for the hypotheses to be tested in the present study.

Graham (1973) suggested that the LPC measures a predisposition to respond behaviorally in certain consistent ways to specific situations, the predisposition being "triggered" or activated by the social environmental factors inherent to those specific situations. This argument seems related to the attitudinal analysis approach of Fishbein et al. (1969) discussed above.

Our thought here is related to both of these pproaches but has a different focus. We uggest that the environment—the situation tself—may moderate not only the LPC–performance relationship but the relationships etween LPC and other psychological meaures as well. This proposition suggests that ither LPC or the other measures (or both) re labile. Such reasoning is quite different rom that proposed by Fiedler (1974), who sserts the long-term stability of LPC and eems relatively unconcerned about relationhips between LPC and any variables other han performance. Our primary hypothesis in the present study is, then, that the relationship between LPC and certain psychological measures is also contingent on the situation.

We can test this hypothesis by examining correlations between LPC and a variety of psychological measures, for a set of leaders, while anticipating no significant results. The sample can then be divided into as many situation types as exist in terms of Fiedler's (1967) definitions, and the correlations recalculated for each subsample. The hypothesis is that the subsamples exhibit markedly different correlations between LPC and the psychological measures used. Examination of the patterns of relationships found in the subsamples might then provide greater understanding of what it is that the LPC instrument measures.

Does LPC predict performance? Since there exists the opportunity, in the above design, for examining relationships between LPC and performance and since there does exist a need for further validation studies to determine more precisely the validity and utility of the contingency model, it seems appropriate to ask this question as well. This is, however, a secondary concern in the present investigation.

METHOD

Setting

The present investigation is one part of a larger study concerned with the psychological and organizational correlates of performance and effectiveness among offices of a nationwide brokerage firm (Taylor, 1972). The total sample consisted of 40 offices in the United States. Data from 34 offices were suitable for the present analysis. Offices are distributed nationally, and the sample reflects this dispersion as well as a bimodal distribution of offices by size. There were 24 offices with staff sizes ranging between 17 and 25 account executives, with a mean of 20, and 16 offices with staff sizes ranging between 26 and 49, with a mean of 30.

Psychological Measures

All measures were self-administered as part of a mail survey package completed by office managers and account executives. Office managers completed the LPC instrument as well as a number of other psychological measures, including: (*a*) a shortened version of the California F Scale (Robinson & Shaver, 1969); (*b*) a shortened version of the Rokeach Dogmatism Scale (Troldahl & Powell, 1965); (*c*) Rotter's (1966) measure of internal-external control attribution; (*d*) the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960); (*e*) Miles's Trust Scale (Miles & Ritchie, 1968); (*f*) Schroder's Sentence Completion Test for Cognitive Complexity (Schroder, Driver, & Streufert, 1967); (*g*) a modification of Scott's (1962) cognitive complexity instrument (Taylor, 1972); (*h*) a measure of openness with information derived from the Taylor (1972) instrument; and (*i*) a desired conflict management style instrument (Taylor, 1972) similar to that developed by Lawrence and Lorsch (1969). Copies of all instruments used can be found in Taylor (1972).

The less familiar Taylor instruments require some brief description. The cognitive complexity measure asks the respondent to list all those securities important to the respondent's work. Each item listed is designated with a letter. The respondent then groups the letters on a separate list, according to common item characteristics, indicating the common characteristic for each group. Space is provided for up to 26 items and up to 12 groups of letters. This procedure was repeated for firm members and for customers. Complexity is defined as the number of items which are assigned unique profiles of characteristics in the grouping process.

The original lists of securities, firm members, and customers could be removed from the questionnaire by the respondent. Since each item was coded with a letter and since the measure requires only the list of groupings, the initial list with items identified was not needed. Respondents were instructed to remove these initial lists if they wished to do so. If the respondent removed the list, this was scored as being "closed with information." Leaving the list in the questionnaire was scored as being "open."

The desired conflict management style instrument uses precisely the methodology developed by Lawrence and Lorsch (1969). Respondents rate each of 30 proverbs on a 5-point scale ranging from 1 (a very undesirable way to handle disagreements or conflicts) to 5 (a highly desirable way). The 30 items were previously shown to form five factors (Taylor, 1972). Individuals are scored on preference for confronting, smoothing, compromising, force, and withdrawal by summing the item scores which comprise each of these factors. Note, then, that the category measures are independent, and an individual respondent could, conceivably, attain any combination of high- and low-preference scores on the five factors.

Situational Measures

The three situational variables of primary significance in the contingency model are (*a*) leader–member relations, (*b*) task structure, and (*c*) leader position power. Items from the Survey of Organizations Questionnaire (Taylor & Bowers, 1970) were used to measure *a* and *c*. We should note that the type of leader–member relations measure used has varied widely in studies of the contingency model, for example, sociometric questionnaires, satisfaction with leader scales, and leaders' or group members' ratings of group atmosphere. Results obtained using leader versus group member ratings have in some cases shown considerable differences (Fiedler, 1967). These results, however, have generally involved leaders' responses on a group atmosphere scale which McMahon (1972) has noted bears considerable resemblance to the LPC measure. To avoid any methodological confounding, the measure used here was a 5-point rating of satisfaction with supervisor, as reported by office members. This item yielded a considerable range of results, office means going from 2.31 (low) to 4.83 (high), with 3.88 being the overall mean of office means. For situational definition, scores below 3.60 were designated low, which indicated poor leader–member relations, while scores of 3.60 and above were designated high, which indicated good leader–member relations. This "split" (at 3.60) was chosen such that the low-satisfaction offices would be certain to fit this category. That is, the overall mean of 3.88 represents, in all likelihood, some degree of upward bias in responses. The 3.60 division thus represents a conservative judgment by the researchers. An extensive empirically based discussion of this issue can be found in Taylor and Bowers (1970; see, particularly, chap. 5).

The task was the same for all subordinates in all offices: the trading of securities for clients. Application, by the authors, of the four criteria used by Fiedler (1967, p. 28), which were developed by Shaw to evaluate the degree of task structure, indicated the task to be relatively unstructured.

Leader position power was measured by office members' responses to the Survey of Organizations Questionnaire item, "How much say or influence does each of the following people have on what goes on in this office?" followed by separate listings of each position by organizational level: top managers, office manager, operations manager, and account executives. Each position was rated on a 5-point scale, ranging from 1 (little or none) to 5 (a very great deal). This item seems to provide a reasonably good measure of the structural or positional distribution of power in an organization (Tannenbaum, 1968). All office managers were reported by subordinates to have relatively high influence, the range being from 3.43 to 5.00, with a mean of office means equal to 4.35.

Fiedler's model calls for a high or low dichotomized catagorization of leader position power. These data indicate that while some office managers are, indeed, relatively higher or lower in position power than others, none could reasonably be considered to be in the low category. It was thus concluded that office manager position power was high in all cases.

The above definitions of the three contingency model situational variables split the 34 offices into two groups, fitting Situation Types 3 or 7 of the model, differing only in leader–member relations: good–Situation Type 3 and poor–Situation Type 7. For the rather favorable situation (Type 3), low-LPC leaders are predicted to be more effective than leaders with high-LPC scores, while for the unfavorable situation (Type 7) the reverse is true. The specific LPC–performance correlations predicted by the model are −.29 and +.26, respectively (Fiedler & Chemers, 1974). We should note, however, that the present study involves groups which fit more closely with the definition of *coacting* groups, "groups whose tasks . . . involve virtually independent performance by the group members, even though there might be some aggregate team score [Fiedler & Chemers, 1974, p. 87]." For such cases, data are sparse for Situation Types 3 and 7 but suggest that the predicted correlations might be even larger (Fiedler, 1971c).

Performance Measures

The measure of leader effectiveness was a "performance ratio," defined as the ratio of the sum of commissions *actually* earned by all account executives in the office to the sum of commissions which *would have* been earned if each account executive had performed according to the average for persons with the same amount of experience in the firm. In addition, a measure of two-year improvement in performance ratio was available. Both measures were taken from organization performance records.

RESULTS

Before examining specific correlational results, we must be certain that differential effects for the two subsamples could not be produced by some selection bias involving the predictor variable, LPC. The mean LPC scores for office managers in the two situations were virtually identical: 61.2 and 62.1 for Situation Types 3 and 7, respectively. We may, therefore, conclude that high- or low-LPC leaders were not, in some biased way, selected into one or the other situation type. A second concern is whether the moderator and predictor variables are related in any way which might bias the predictor–moder-

TABLE 1

CORRELATIONS BETWEEN LEAST PREFERRED CO-WORKER AND OTHER MEASURES AS MODERATED BY SITUATIONAL FACTORS

Other measure	Correlation between LPC and other measure			
	All offices ($n = 34$)	Situation Type 3 ($n = 22$)	Situation Type 7 ($n = 12$)	Absolute difference (3 versus 7)
Authoritarianism	+.12	+.13	+.24	.11
Dogmatism	+.01	+.09	−.12	.21
Defensiveness (Marlowe-Crowne)	+.00	−.05	+.16	.21
Internal control	+.17	+.14	+.35	.21
Trust in subordinates	+.09	+.08	+.18	.10
Cognitive complexity				
Sentence Completion Test	−.02	+.04	−.16	.20
Classifying test (Taylor)				
Securities (nonsocial)	+.18	+.45	−.03	.48
Firm members (social)	+.40*	+.63**	−.32	.95*
Customers (social)	+.31	+.37	−.20	.57
Openness with information				
Securities	−.18	−.46*	+.55*	1.01**
Firm members	−.09	−.34	+.74*	1.08**
Customers	−.23	−.42	+.85*	1.27**
Desired conflict management style				
Confronting resolution	+.14	+.03	+.43	.40
Smoothing resolution	+.13	−.02	+.55*	.57
Compromising	+.03	−.18	+.70**	.88**
Force	−.34	−.30	−.48	.18
Withdrawal	+.21	+.21	+.24	.03
Satisfaction with office manager	−.01	−.01	+.57*	.58
Performance				
Current year	+.09	+.07	+.18	.11
Two-year improvement	+.10	−.05	+.34	.39
Other studies[a]				
Interacting groups		−.29	+.26	.59
Coacting task group[b]		−.32	+.87	1.19

Note. LPC = Least preferred co-worker.
[a] Data taken from Fiedler and Chemers (1974).
[b] $N = 1$.
* $p < .05$.
** $p < .01$.

ator-outcome variable relationships. Again, such an effect can be reasonably ruled out, since the correlation between office managers' LPC, the predictor, and office members' satisfaction with leader, the moderator, was essentially zero for the 34 offices ($r = +.005$, *ns*). Having excluded the possibility of such sampling bias effects, we may proceed to examine our results for the hypothesized moderator effect.

Situational Moderator Effects

Table 1 presents the correlations between LPC scores and the scores on the various psychological measures used, for the total sample of offices and for the two subsamples. As can be seen, none of the measures showed any overall relationship with LPC except for cognitive complexity, results conforming to those obtained by Mitchell (1970), and the preference for a forcing conflict management style by low-LPC managers. Of these results, only social cognitive complexity as measured by categorization profiles of firm members was statistically significant.

The correlations for each subsample, however, show a pattern of significantly different relationships for some of the measures. While only 7 of the 34 correlations were, of themselves, significant, the best direct test of our hypothesis, which predicts significantly different correlations between LPC and the other psychological measures for the two subsamples, is a comparison of the pairs of correlations. Results showed that these differences between pairs were significant in 5 of 17 cases. Of the 12 remaining pairs of correlations, our hypothesis was directionally confirmed in six cases, in that the correlations for Situation Types 3 and 7 were signed oppositely. Overall, then, for 11 of 17 measures, we found opposite, and in 5 cases, significantly different, correlations between LPC and the particular measure in the two situations.

We should note that two of the measures used, authoritarianism and dogmatism, are short form versions of the standard instrument, and might be considered inadequately validated. In two other cases, internal control and cognitive complexity/nonsocial, an analysis of various population statistics for the two subsamples showed these two situation type groups to be nonequivalent. Thus, these comparisons are somewhat questionable. Neither of these factors, however, seriously affects the basic outcome of the present analysis. We could say that of 13 comparisons, dropping the 4 in question, our hypothesis was supported in 9 cases directionally, and the opposite-signed correlations were significantly different in 5 of the 13 cases.

Clearly, the strongest support for the hypothesized moderating effect came from three measures: (*a*) openness with information, (*b*) cognitive complexity, and (*c*) preferred conflict management styles. We can see that on all three openness measures, greater openness was associated with *lower* LPC scores in Situation Type 3, while the reverse was true in Situation Type 7. For the social measures, greater cognitive complexity was associated with higher LPC scores in Situation Type 3 but with lower LPC in Situation Type 7. Desired conflict management style showed little relationship to LPC in Situation Type 3, with the exception of a tendency for high-LPC managers to avoid expressing a preference for force and an even weaker tendency to avoid expressing a preference for compromise. However, for managers in Situation Type 7, LPC was rather clearly related to expressed preference of conflict management style. Specifically, in Situation Type 7 environments, high-LPC office managers perferred compromise, smoothing resolution, and confronting resolution and preferred to avoid force.

We should also note one unexpected result: Satisfaction with office manager, our moderating variable, showed no relationship to LPC either overall or in Situation Type 3 but was positively and significantly correlated with LPC for the 12 Situation Type 7 offices.

Least Preferred Co-worker Score and Performance

Our secondary concern was for a test of the hypothesis generally investigated by Fiedler and his associates concerning the relationship between LPC of the leader and group performance, for the two situations identified earlier. Fiedler's recent summary analyses (1971c; Fiedler & Chemers, 1974) showed a median correlation for interacting groups in field studies of −.29 for Situation Type 3 and +.26 for Situation Type 7. A single study of coacting task groups yielded correlations of −.32 for Situation Type 3 and +.87 for Situation Type 7. The correlations obtained in the present analysis, using performance ratio as the outcome variable, were +.07 and +.18 for Situation Types 3 and 7, respectively. Neither of these results attained or approached statistical significance, although the result for Situation Type 7 was in the predicted direction and reasonably approximated the outcome predicted on the basis of field studies ($r = +.26$).

When performance was defined as improvement over a two-year period, the results were even closer to the predictions of the contingency model: −.05 for Situation Type 3 and +.34 for Situation Type 7. While, again, not statistically significant, the directional conformance with Fiedler's reported results was greater in both cases, and the magnitude of the correlation for Situation Type 7, +.34, versus a prediction within the range of from +.26 to +.87, provides some support for the predictive validity of the contingency model.

DISCUSSION AND CONCLUSIONS

We are now able to attempt to answer the two questions postulated for this study: (*a*)

What does the LPC instrument measure? and (*b*) Does LPC predict performance in the situations studied?

We have reviewed and summarized Fiedler's beliefs about what the LPC measures (Fiedler, 1971a; Fiedler & Chemers, 1974) and noted some alternate explanations (e.g., Fishbein et al., 1969; Graham, 1973). Our results suggest a rather different interpretation from any proposed to date.

The data in Table 1 suggest rather strongly that our hypothesis—situational variables moderate the relationship between LPC and other psychological measures—is at least on target, if not fully confirmed. Most consistent in exemplifying the present hypothesis were the information openness results; all three measures, which showed no initially significant relationship with LPC, yielded significant differences when the situation was taken into account. The cognitive complexity measures followed a similar pattern, as did conflict management preference, with six of these nine measures yielding differential results consistent with the hypothesis.

In conclusion, it has been shown that situational factors do, in some cases, moderate the relationship between LPC and other psychological measures. Before considering the implications of this finding, we might briefly speculate as to why our results were not even stronger and more consistent.

Perhaps the most obvious factor is the situational moderator used, office members' satisfaction with the office manager, which reflects leader–member relations. While the nature of leader–member relations has been repeatedly shown to be a significant organizational variable (Argyris, 1957; Bowers & Seashore, 1966; Kahn, 1958; Likert, 1961; Morse & Reimer, 1956), it is surely not the only such factor and probably not always the most important of potential situational moderators. In other words, the larger issue at this point seems to be the determination of what the most significant moderating variables are. This same point has been made by Fiedler and Chemers (1974) with respect to situational measures for best defining the contingency model situation types. Assuming that further research results in some general list of these moderating factors, let us consider the implications of our results for the interpretation of the LPC measure.

We have already concluded that LPC does, in certain situations and to some degree, predict leader effectiveness in terms of group performance (Fiedler & Chemers, 1974; Sashkin, 1972). The present study adds some slight weight to such a conclusion, at least for Situation Type 7. The confirmation is greater when performance improvement is considered rather than current performance. Thus, the situation moderates the relationship between LPC and performance. Given the fact that LPC is, it would seem, uniquely unrelated to any other psychological measure (except perhaps cognitive complexity), it is reasonable, and perhaps even unsurprising, to find that the situation can also moderate the relationships between LPC and other psychological measures.

Since the LPC–performance relationship is strongest for Situation Type 7, let us first examine, in more detail, the relationships between LPC and other psychological measures for this situation. The high-LPC leaders have been shown to be somewhat more effective than the low-LPC leaders. They are also more open with information and less socially differentiating. As compared with low-LPC leaders in this situation, high-LPC office managers prefer relatively noncoercive conflict management styles—compromise, smoothing resolution, and confronting resolution—over force. Finally, although Situation Type 7 was defined by low satisfaction with the office manager, we see that even within this obviously narrowed range of satisfaction scores (2.31 to 3.59) there is a strong positive relationship between satisfaction and office managers' LPC scores. These results suggest that the high-LPC office manager in this moderately unfavorable situation is, as Fiedler (1971a) asserted, behaviorally more relationship oriented than the low-LPC leader.

In looking at Situation Type 3, we might naturally expect to find just the reverse pattern, since high-LPC leaders were predicted to be more effective than low-LPC office managers. Recall, however, that this was not the case. In this study, the correlation between performance and LPC in Type 3 situations was essentially zero. Note, then, that while the pattern of relationships between LPC and the other measures (see Table 1) for Situation Type 3 is reversed for openness and cognitive complexity, this is not the case for the conflict management preference measures. In Type 3 situations, low-LPC office managers did not prefer smoothing or confronting to any greater extent than did high-LPC leaders and showed only a slight tendency to prefer compromise. Furthermore, low-LPC leaders still showed greater preference for forcing in Type 3 situations. And, finally, we see that in this situation there was *no* relationship between LPC and satisfaction with the leader. In conclusion, the fact that low-LPC office managers in this situation (Type 3) were not superior to high-LPC leaders could possibly be due to an inadequate relationship orientation.

What, then, does the LPC score mean? We cannot definitely answer this question, given the present data. Our results would indicate, as was earlier suggested, that some of the measures employed are labile. Whether this means that LPC can change over situations or that the other psychological measures exhibit such shifts is a question for further research.[5] It does, however, seem that our results support rather strongly the inference that effectiveness is related to behavioral differences among leaders which, in certain circumstances, are very similar to the expectations of leader behavior which can be derived from the contingency model (Fiedler, 1971a; Fiedler & Chemers, 1974).

In conclusion, we would suggest that LPC may have no meaning outside of situational contingencies. Leaders' LPC score—the degree of esteem reported for some particular least preferred co-worker—may have a more general meaning only in the context of actual situations. It is in these specific, defined situations that LPC is a potentially critical and useful link between the leader, as defined by psychological and behavioral measures, and the performance of the group.

---

[5] A recent study by Stinson and Tracy (1974) showed far greater instability in LPC scores than had been assumed on the basis of earlier research (Fiedler, 1967).

REFERENCES

Argyris, C. *Personality and organization*. New York: Harper & Row, 1957.

Bowers, D. G., & Seashore, S. E. Predicting organizational effectiveness with a four-factor theory of leadership. *Administrative Science Quarterly*, 1966, **11**, 238–263.

Chemers, M. M., & Skrzypek, G. J. An experimental test of the contingency model of leadership effectiveness. *Journal of Personality and Social Psychology*, 1972, **24**, 172–177.

Crowne, D. P., & Marlowe, D. A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 1960, **24**, 349–354.

Duffy, J. F., Kavanagh, M. J., MacKinney, A. C., Wolins, L., & Lyons, T. F. A field extension of Fiedler's contingency model. Paper presented at the 42nd Annual Meeting of the Midwestern Psychological Association, Cincinnati, Ohio, April 1970.

Fiedler, F. E. *A theory of leadership effectiveness*. New York: McGraw-Hill, 1967.

Fiedler, F. E. *Leadership*. New York: General Learning Press, 1971. (a)

Fiedler, F. E. Note on the methodology of the Graen, Orris, and Alvares studies testing the contingency model. *Journal of Applied Psychology*, 1971, **55**, 202–204. (b)

Fiedler, F. E. Validation and extension of the contingency model of leadership effectiveness: A review of empirical findings. *Psychological Bulletin*, 1971, **76**, 128–148. (c)

Fiedler, F. E. Personality, motivational systems, and behavior of high and low LPC persons. *Human Relations*, 1972, **25**, 391–412.

Fiedler, F. E., & Chemers, M. M. *Leadership and effective management*. Glenview, Ill.: Scott, Foresman, 1974.

Fishbein, M., Landy, E., & Hatch, G. Some determinants of an individual's esteem for his least preferred co-workers: An attitudinal analysis. *Human Relations*, 1969, **22**, 173–188.

Graen, G., Alvares, K. M., Orris, J. B., & Martella, J. A. Contingency model of leadership effectiveness: Antecedent and evidential results. *Psychological Bulletin*, 1970, **74**, 285–296.

Graen, G., Orris, J. B., & Alvares, K. M. Contingency model of leadership effectiveness: Some experimental results. *Journal of Applied Psychology*, 1971, **55**, 196–201.

Graham, W. K. Description of leader behavior and evaluation of leaders as a function of LPC. *Personnel Psychology*, 1968, **21**, 457–464.

Graham, W. K. Leader behavior, esteem for the least preferred co-worker, and group performance. *Journal of Social Psychology*, 1973, **90**, 59–66.

Kahn, R. L. Human relations on the shop floor. In E. M. Hugh-Jones (Ed.), *Human relations and modern management*. Amsterdam: North-Holland, 1958.

Larson, L. L., & Rowland, K. M. Leadership style and cognitive complexity. *Academy of Management Journal*, 1974, **17**, 37–45.

Lawrence, P. R., & Lorsch, J. W. *Organization and environment*. Homewood, Ill.: Irwin, 1969.

Likert, R. *New patterns of management*. New York: McGraw-Hill, 1961.

McMahon, J. T. The contingency theory: Logic and method revisited. *Personnel Psychology*, 1972, **25**, 697–710.

Miles, R. E., & Ritchie, J. B. Leadership attitudes among union officials. *Industrial Relations*, 1968, **8**, 108–117.

Mitchell, T. R. Leader complexity and leadership style. *Journal of Personality and Social Psychology*, 1970, **16**, 166–174.

Morse, N. C., & Reimer, E. M. The experimental change of a major organizational variable. *Journal of Abnormal and Social Psychology*, 1956, **51**, 120–129.

Rokeach, M. *The open and closed mind*. New York: Basic Books, 1960.

Rotter, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 1966, **80**(1, Whole No. 609)

Robinson, J. P., & Shaver, P. R. *Measures of social psychological attitudes*. Ann Arbor, Mich.: University of Michigan, Institute for Social Research, Survey Research Center, 1969.

Sample, J. A., & Wilson, T. R. Leader behavior, group productivity, and rating of least preferred co-workers. *Journal of Personality and Social Psychology*, 1965, **1**, 266–270.

Sashkin, M. *Supervisory leadership in problem solving groups: Experimental tests of Fred Fiedler's "Theory of Leadership Effectiveness" in the laboratory using role play methods*. (Doctoral dissertation, University of Michigan, 1970) Ann Arbor, Mich.: University Microfilms, 1971, No. 71-4719.

Sashkin, M. Leadership style and group decision effectiveness: Correlational and behavioral tests of Fiedler's contingency model. *Organizational Behavior and Human Performance*, 1972, **8**, 347–362.

Schroder, H. M., Driver, M. J., & Streufert, S. *Human information processing*. New York: Holt, Rinehart & Winston, 1967.

Scott, W. A. Cognitive complexity and cognitive flexibility. *Sociometry*, 1962, **25**, 405–414.

Stinson, J. E., & Tracy, L. Some disturbing characteristics of the LPC score. *Personnel Psychology*, 1974, **27**, 447–485.

Tannenbaum, A. S. *Control in organizations*. New York: McGraw-Hill, 1968.

Taylor, F. C. *Associations among personal and organizational characteristics in forty branch offices of a securities brokerage firm*. (Doctoral dissertation, University of Michigan, 1971) Ann Arbor, Mich.: University Microfilms, 1972, No. 72-4991.

Taylor, J. C., & Bowers, D. G. *Survey of organizations: A machine-scored standardized questionnaire instrument*. Ann Arbor, Mich.: University of Michigan, Institute for Social Research, Center for Research on Utilization of Scientific Knowledge, 1970.

Troldahl, V. C., & Powell, F. A. A short-form dogmatism scale for use in field studies. *Social Forces*, 1965, **44**, 211–214.

(Received December 28, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 6, 741-747

SOME RESULTS WITH THE GERMAN FORM OF THE POSITION ANALYSIS QUESTIONNAIRE (PAQ)[1]

E. FRIELING,[2] W. KANNHEISER, AND R. LINDBERG

University of Regensburg, Regensburg, Germany

This study involved the development of a German form of the Position Analysis Questionnaire (PAQ) to serve as a structured job analysis procedure. In Phase 1, independent translations of the PAQ, made by three groups of psychologists, were incorporated into the PAQ Bd (the German edition). Thirty trained graduate students analyzed 61 positions with the PAQ Bd. Average interrater reliability was .79; average reliability of the individual job elements was .65. In Phase 2, 70 industrial psychologists rated the importance of 59 attributes, modified to include better definitions, for each of the 205 job elements. Satisfactory levels of reliability were achieved. It seems possible to construct profiles of job requirements for positions on the basis of the PAQ Bd analyses of jobs in combination with the attribute profiles of the individual elements.

Data obtained through job analysis procedures are useful for several functions related to personnel management, such as in the establishment of personnel specifications, training, job classification including the establishment of wage and salary rates, and job design. Conventional essay types of job descriptions can well serve useful purposes in connection with such functions. However, for some personnel management functions and for research purposes supporting such functions, some types of standardized job analysis procedures seem to be particularly desirable. Such a procedure preferably should fulfill certain conditions; it should (*a*) have a sound theoretical base, with a clear definition of the nature of the analysis; (*b*) provide a standardized approach applicable to a wide range of jobs; (*c*) permit statistical quantification of the data in terms of relevant units of job data; (*d*) lend itself to reliable use; and (*e*) be economically feasible.

A review of the relevant German literature revealed no such job analysis procedure which might meet these conditions for use in the Federal Republic of Germany. However, the Position Analysis Questionnaire (PAQ), developed in the United States by McCormick, Jeanneret, and Mecham (1972), seems to meet certain of these conditions. The PAQ is a highly structured type of job analysis questionnaire, consisting of nearly 200 job elements (or items) of a "worker-oriented" nature (McCormick, 1959; McCormick, Cunningham, & Thornton, 1967), which yields information about job activities, their corresponding working conditions, and the specific situations which might influence the activities of workers.[3]

The present study consisted of two phases:

1. The translation and transformation of the PAQ into an adequate German form and the use of that form in the analysis of a sample of jobs, to evaluate its relevance as a job analysis instrument in the context of the Federal Republic of Germany.
2. The development of a possible data base of ratings of human attributes required for the job elements of the PAQ that might be used in determining the psychological and physical requirements for jobs.

---

[1] The initial research with the Position Analysis Questionnaire (PAQ) was reported in a monograph by McCormick, Jeanneret, and Mecham (1972).

The authors of this study wanted to find out whether it would be possible to use a German form of the PAQ under German working conditions. The study was financed by the Bundesanstalt für Arbeit (Federal Republic of Germany).

[2] Requests for reprints should be sent to Ekkehard Frieling who is now at Memminger Strasse 14, University of Augsburg, 8900 Augsburg, Germany.

[3] Sample job elements (without their more detailed descriptions) include: use of written materials, reasoning in problem solving, hand–arm manipulation, etc.

PHASE 1: INITIAL DEVELOPMENT AND EVALUATION

German Form of the PAQ

The initial research with the PAQ reported by McCormick et al. (1972) was based on Form A. Form B was used as the basis of the translation. Although reliability data are not available for Form B, this form presumably incorporates a number of improvements over Form A. The German edition is referred to as PAQ Bd. Because of linguistic, educational, and cultural differences between the United States of America and the Federal Republic of Germany, the following modifications seemed desirable: (*a*) Certain modifications of the instructions. (*b*) Modification of certain rating scales. Two of the six types of rating scales used in rating the job elements ("Importance" and "Extent of Use") had to be translated rather freely to provide subjectively equal distances between the steps of the scales. (*c*) Modification of certain job elements (or items). Because of cultural differences certain job elements (46, 47, and 48) were materially altered. (*d*) Addition of new items. On the basis of pilot studies, it seemed advisable to add certain new job elements, such as "knowledge of foreign languages" (which is an important requirement for many jobs in Germany). (*e*) Elimination of job elements. The "income/pay" items were eliminated, along with four "open-ended" items that required write-in responses.

On the basis of these modifications, three different groups of psychologists from the "Bundesanstalt für Arbeit" in Nürnberg, Germany, the "Forschungsinstitut für praktische Psychologie" in Salzburg, Austria, and the first author (University of Augsburg) translated the PAQ independently of the other groups. The resulting translations were then discussed and incorporated into the PAQ Bd.[4]

Sample of Positions Analyzed

Sixty-one positions were analyzed with the PAQ Bd. According to McCormick, very heterogenous samples of positions should be used in the computation of the reliability of the PAQ in order to increase the probability of inclusion of jobs that represent all scale values of the various job elements. In our numerically limited sample this was not possible in all cases. Nevertheless, efforts were made to include as many different types of positions as possible, including, for example, those from different areas of the electrical industry, the Federal Railway, the Federal Post, the chemical industry, the glass working industry, and several trades. As it turned out, the ratings on most job elements did spread out over the entire range of scale values from 0 to 5.

Analysts and Analysis Procedures

Thirty analysts, randomly combined in pairs, analyzed these 61 positions. Based on the scheme by Heyns and Zander (1953), a training course for these analysts was developed. The sample consisted of graduate students of industrial psychology at the University of Regensburg. The analysis procedure was the same as the one used by McCormick, Jeanneret, and Mecham (1969b). Specifically for most of the job elements there was an appropriate scale such as "Importance", "Extent of Use", "Time", etc. or a special scale applicable to the specific element; the analyst rated each such element as it related to a given position or job, using in most instances a 6-point scale from 0 (does not apply) to 5 (the highest scale value). Several elements were of a dichotomous nature; in such instances the analyst marked the element as "0" (does not apply) or "1" (does apply). Each analyst was paid the equivalent of about $12 (American).

Reliability of Analyses

In order to be able to evaluate the reliability of the analyses, two analysts analyzed each position. The correlation of the ratings of these pairs of analysts would provide information about their consistency. With this procedure it would be possible to compare our data with reliability data based on PAQ Form A, as reported by McCormick et al. (1969b).

One reliability analysis consisted of correlating the ratings of the job elements as made by the two analysts of each pair who analyzed a given position. The correlation for each position was then made for the two ratings across all job elements.

---

[4] This translation of the PAQ Form B (McCormick, Jeanneret, & Mecham, 1969a) was authorized by the Purdue Research Foundation (Nov. 1970).

Since 61 positions were analyzed by pairs of independent analysts, it was also possible to determine the reliability of the ratings for each job element. This was done by computing a correlation coefficient for each job element, giving a total of 207 coefficients. Because the responses to certain items were of essentially a *nominal* nature, the reliability of these items was determined with the "Kappa coefficient" developed by Cohen (1960).

Results

The reliability coefficients for the individual pairs of analysts ranged from .51 to .93, with an average reliability of .79 (computed across *z*-transformed correlations). In the case of the reliability coefficients of the individual job elements, the average coefficient was .65. Half of the coefficients were higher than .61; the coefficients for 23 job elements were insignificant (i.e., less than .35).

Discussion

The average reliability coefficients of the pairs of raters (.79) were identical with those reported by McCormick et al. (1969b). With respect to the reliability coefficients of the individual job elements, McCormick et al. found an average coefficient of .80 which is considerably higher than ours (.65).

Although our interrater reliability coefficients corresponded with those reported by McCormick et al. (1969b), there is some question as to why the reliability coefficients of the individual job elements in our study were somewhat lower than those reported by McCormick. It is of course possible that these differences are due to differences between PAQ Form A and Form B, although in McCormick's opinion Form B should be at least as reliable as Form A. On the other hand, the PAQ Bd might not be as reliable as the PAQ Form A because of the translation. In addition there might be intercultural differences with respect to the content of several terms included in the American or German items. However, the results of a later pilot study covering 32 positions indicated that certain modifications in the content and form of those low-reliability job elements increased overall reliability. In this study, the average reliability of the job elements was .78. Although the size of the sample used in this pilot study had to be limited, it seems probable that the reliability of the job elements might be increased further by additional modifications and the use of a larger sample.

In general terms, the trial use of the PAQ Bd was sufficiently encouraging to suggest that it might well serve as a structured job analysis procedure for various purposes, especially since it appears that some further modification of selected items might enhance its reliability in use.

PHASE 2: DEVELOPMENT OF DATA BASE OF ATTRIBUTE PROFILES OF JOB ELEMENTS

The results of the first phase encouraged the investigators to attempt to derive estimates of the *requirements* (in terms of certain psychological and physical attributes) of the individual job elements of the PAQ Bd. The method suggested by McCormick et al. (McCormick, 1959; McCormick, Cunningham, & Thornton, 1967; McCormick, Jeanneret, & Mecham, 1969b; Mecham & McCormick, 1969) seemed likely to serve as an appropriate procedure for obtaining estimates of the attributes required for the approximately 200 individual job elements. This method provides for experts to rate the relevance of each of certain attributes to the individual job elements (see Mecham, 1968; Mecham & McCormick, 1969). In contrast to the procedures followed by Theologus, Romashko, and Fleishman (1970) and Theologus and Fleishman (1971), who sought to determine job requirements for relatively complex tasks through use of task anchored ability rating scales, McCormick's concept divides positions and tasks into "elements" for which requirements are estimated. Since it is believed that these elements are easily grasped and relatively unequivocal, it may be supposed that overly global ratings might be avoided. The ratings may be combined with the results of the PAQ analyses of a given position in order to construct or build up a profile of the total job requirements for that position. This profile is based on the consolidated requirements of the job elements that comprise the position. This procedure presumably would be quite economical.

The following part of this study describes the selection and rating of the attributes as related to the job elements of the PAQ Bd.

Selection of Attributes[5]

The list of attributes used in this phase was one developed by Wenninger, Kannheiser, Frieling, and Hoyos (1971), which contained 59 physical and psychological attributes. In developing this list, an extensive analysis of the relevant literature had been carried out. More than 400 sources were analyzed for information about definitions of the attributes and attribute measuring devices; most of these sources were factor-analytic studies concerned with the analysis of simple and complex perception, psychomotor skills, physical abilities, intelligence, temperament traits, and interests.

The selection of the attributes for use in the second phase of the study was carried out with special consideration of the following points which seemed extremely important: (*a*) Attributes taken into the list had to have positive correlations with successful performance in different jobs, as determined in previous studies. (*b*) To meet the criterion of "measurability" and "methodological clearness", only attributes based on the results of factor-analytic studies were taken into account. (*c*) The authors also tried to obtain a high degree of applicability of the attributes by defining them as clearly and precisely as possible. The attributes were defined in relatively close relationship to the primary literature. In contrast to the attributes of the list proposed by Mecham (1968), special explanations and measurement instruments similar to those provided by Theologus, Romashko, and Fleishman (1970) were given.

A comparison should be made of the 59 attributes used in this study as compared with the 68 used by Mecham (McCormick, Jeanneret, & Mecham, 1972; Mecham, 1968). Mecham's attributes were in two classes, namely those of an *aptitude* nature (including the types of qualities normally considered as aptitudes) and those of a *situational* nature (those characterizing situations to which workers need to adjust, such as working alone or adapting to repetitive operations). Of the 59 attributes used in the present study, 47 corresponded essentially with those used by Mecham, although the definitions of certain of these were modified. Seven of Mecham's aptitudes were omitted, but five others were added.[6] Most of the differences, however, related to Mecham's situational attributes. Since these presumably have implications in terms of personality characteristics (in the sense of the worker having to adjust to the work situation), some of these were replaced with more specific temperament and interest traits. It should be noted that in the comparison of the two sets of attributes, in a limited number of cases one attribute in one set was considered to embrace two in the other set, and vice versa. Therefore a strict one-to-one comparison is a bit misleading.

Selection of Raters

One hundred and thirty-five industrial psychologists were chosen at random from the membership list of the German Association of Industrial Psychologists. Each of these psychologists was asked to participate as a rater in our research project. Of 135 psychologists, 76 psychologists were willing to cooperate; 27 refused to cooperate; 31 did not answer at all. From the 76 cooperative psychologists, 70 were randomly selected as raters. Of these raters, 37% worked as personnel managers, 20% taught industrial psychology in German universities, and 16% were employed in departments for the selection and placement of industrial personnel. The rest (18%) of the raters worked in different areas of industrial psychology and related fields. Each rater was paid the equivalent of about $100 (American).

---

[5] Sample attributes include: perceptual attributes (spatial orientation, flexibility of closure), physical abilities (body strength), memory (visual memory), etc.

[6] Specific changes may be obtained from the authors.

Rating Procedures

To be able to carry out the ratings, each rater was provided with the following rating materials: (*a*) a general introduction to the purpose of the study; (*b*) detailed instructions; (*c*) a copy of the modified PAQ; (*d*) 10 rating manuals with detailed definitions of the attributes which were supposed to be rated and several explanations; and (*e*) a short inventory for the recording of socioeconomic data.

Each rater had to rate the importance of each of 10 attributes with respect to the 205 job elements on a 6-point scale ranging from 1 (lowest importance) to 6 (highest importance). This rating scale seemed an adequate translation of the one used by Mecham (1968, p. [illegible]9).

Suitable action was undertaken to restrict the influence of systematic errors in the rating process as much as possible by the following procedures: (*a*) The 70 groups of attribute materials (10 attributes per group to be rated by each of the 70 raters) were assigned to raters on the basis of a latin square design, to eliminate the effects of sequence. (*b*) Attributes that seemed to be logically related to each other were always rated by different raters to minimize the *proximity* error. (*c*) To avoid the *logical* error, definitions and explanations about the attributes were described in detail to provide the raters with extensive knowledge about each attribute. IBM Type 551 data sheets were used to record the ratings. Each rater had to finish his ratings within two months.

Results

On the basis of the medians of the ratings, an "attribute profile" for each of the 205 job elements of the PAQ Bd was computed. From these profiles, it could be deduced how important each of the 59 attributes had been judged for each job element. The medians were divided into two groups:

1. Median < 3.5: the attribute is not important for the job element.
2. Median ≥ 3.5: the attribute is important for the job element.

By summarizing the medians which were equal or greater than 3.5 for each attribute, a rough measure of the judged general importance of this attribute for each of the job elements of the PAQ Bd was available.

To determine the reliability of the ratings, three methods were used: (*a*) The reliability of ratings for each attribute as related to each job element was computed on the basis of the method proposed by Finn (1970). These values may be regarded as measures of the reliabilities of the arithmetic means (and, on the assumption of a high positive correlation between arithmetic means and medians, as a reliability estimation for the medians). (*b*) The reliability with which all the raters (the number varied from 9 to 14) independently rated the attributes across all job elements was computed on the basis of the intraclass correlation (Ebel, 1951; Winer, 1962). (*c*) The reliability of attribute ratings of job elements was computed for each job element across all attributes. Again the intraclass correlation method was used. These coefficients were based on anywhere from 9 to 14 ratings, in most cases 10, 11, or 12.

The complete matrix of medians and the other results are given in Frieling, Kannheiser, and Hoyos (1973). The coefficients of reliability (by the method proposed by Finn) were all significant on the .05 level. The reliability coefficients of the attributes are summarized in Table 1. That table shows the numbers of coefficients in each of several class intervals, including those which generally corresponded to attributes used by Mecham as well as the *new* attributes in the German version and of all those in the German version. For comparative purposes, the coefficients obtained by Mecham are also given (in the first column). It can be seen that for the German version, 11 of the coefficients were .95 or above, 31 were between .90 and .94, 13 from .85 to .89, and 4 from .80 to .84.

Discussion

The attribute ratings generally reflect a satisfactory level of reliability, which is a prerequisite for their potential use as a basis for the construction of requirement profiles for individual jobs. They generally are compa-

TABLE 1

SUMMARY OF RELIABILITY COEFFICIENTS OF ATTRIBUTE RATINGS OF JOB ELEMENTS IN GERMAN VERSION OF THE POSITION ANALYSIS QUESTIONNAIRE AND OF CORRESPONDING ELEMENTS IN THE UNITED STATES VERSION

Class interval of coefficients	Attributes common to both versions		German version	
	United States coefficients[a]	German coefficients	New attributes	All attributes
.95[a]	15	9	2	11
.90 to .94	16	24	7	31
.85 to .89	11	10	3	13
.80 to .84	2	4		4
.75 to .79				
.70 to .74	1			
.65 to .69	2			
Total	47	47	12	59

[a] Data from Mecham (1968) and McCormick, Jeanneret, and Mecham (1972).

rable to those reported by Mecham. A comparison of the coefficients of the *new* individual attributes which deal with temperament and interest characteristics (as contrasted with some of the *situational* attributes of Mecham's that were not included) may be interpreted as a justification of our deviation from Mecham's procedure.

The matrix of median attribute ratings (which can be thought of as reflecting the attribute profiles of the individual job elements) is not given in the present study. An examination of it, however, shows that all attributes contained in the list are relevant for at least some of the job elements. Thus it can be concluded that the list of attributes includes factors which may be relevant requirements for the job elements of the PAQ. The definitions of attributes used seem to explain the function of each attribute so clearly that it was possible to avoid excessively global ratings.

CONCLUSIONS

The two phases of this study have shown that despite the several modifications made in the PAQ and in the attributes used with it, the PAQ Bd yields results which are generally comparable to the ones reported by McCormick et al. (1969b). The reliability coefficients, for the most part fulfilled our expectations. The results of a new pilot study indicate that unsatisfactory reliability coefficients of the use of certain job elements can be increased considerably by modifying these items. The results of the second phase seem to allow several tentative conclusions: (*a*) Experts are capable of rating the importance of human attributes (modified in accordance with procedures such as those designed by Theologus and Fleishman, 1971) for elemental aspects of working behavior with acceptable reliability without being in a real working situation. (*b*) The rated importance of single attributes, with respect to different job elements, varies.

On the basis of these attribute ratings (as reflected in the attribute profiles of the individual job elements of the PAQ Bd), it seems possible to construct plausible profiles of job requirements for different individual jobs. This assumption may be strengthened in several studies which are to be published elsewhere. In these studies it was possible to construct profiles of job requirements for masons, signalmen (railway), and electricians.

The applicability of the PAQ Bd to problems of occupational classification has yet to be studied intensively. At this point we would expect that the PAQ Bd is applicable to these problems, especially as it seems possible to continue constructing profiles of job requirements for positions on the basis of PAQ Bd analyses of jobs in combination with the attribute profiles of the individual elements of it. Thus a psychologically oriented occupational classification may be developed.

In summary, the dimensions of working activities in Germany seem to be comparable to the ones existing in the United States of America, since it is basically possible to use an American job analysis procedure successfully in Germany. Cultural-specific influences seem to be minimal in this case. In general, it seems that occupational classifications based on instruments like the PAQ or the PAQ Bd might well serve as the basis for generating usable statements about similarities of jobs in different countries.

REFERENCES

Cohen, J. A coefficient of agreement for nominal scales. *Educational Psychological Measurement*, 1960, **20**, 37–46.

Ebel, R. L. Estimation of the reliability of ratings. *Psychometrika*, 1951, **26**, 405–433.

Finn, R. H. A note on estimating the reliability of categorial data. *Educational Psychological Measurement*, 1970, **30**, 71–76.

Frieling, E., Kannheiser, W., & Hoyos, C. G. *Untersuchungen zum Position Analysis Questionnaire (PAQ)—Voruntersuchung zur Erstellung von beruflichen Eignungsprofilen*. Erstellt im Auftrag der Bundesanstalt für Arbeit. Regensburg, Germany: 1973.

Heyns, R. W., & Zander, A. F. Observation of group behavior. In L. Festinger & D. Katz (Eds.), *Research methods in behavioral sciences*. New York: Holt, Rinehart, & Winston, 1953.

McCormick, E. J. The development of processes for indirect or synthetic validity. III. Application of job analysis to indirect validity. *Personnel Psychology*, 1959, **12**, 402–413.

McCormick, E. J., Cunningham, J. W., & Thornton, G. C. The prediction of job requirements by a structured job analysis procedure. *Personnel Psychology*, 1967, **20**, 431–440.

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. The development and background of the Position Analysis Questionnaire (PAQ). West Lafayette, Ind.: Purdue University, Occupational Research Center, 1969. (NTIS No. AD-691 737) (Prepared for Office of Naval Research under contract Nonr-1100 (28), Report 5) (a)

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire. Unpublished manuscript, Lafayette, Ind.: Purdue University, Occupational Research Center, 1969. (Prepared for Office of Naval Research under contract Nonr-1100 (28), Report 6) (b)

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire (PAQ). *Journal of Applied Psychology*, 1972, **56**, 347–368.

Mecham, R. C. Ratings of attribute requirements of job elements in a structured job analysis format. Unpublished master's thesis. Purdue University, 1968.

Mecham, R. C., & McCormick, E. J. The rated attribute requirements of job elements in the Position Analysis Questionnaire. West Lafayette, Ind.: Purdue University, 1969.

Theologus, G. C., & Fleishman, E. A. Development of a taxonomy of human performance: Validation study of ability scales for classifying human tasks. Washington, D.C.: American Institutes for Research, 1971.

Theologus, G. C., Romashko, T., & Fleishman, E. A. Development of a taxonomy of human performance: A feasibility study of ability dimensions for classifying human tasks. Washington, D.C.: American Institutes for Research, 1970.

Wenninger, G., Kannheiser, W., Frieling, E., & Hoyos, C. G. *Systematische Zusammenstellung von Attributen für die Beschreibung von psychischen und physischen Anforderungen innerhalb des Arbeitsgeschehens*. Erstellt im Auftrag der Bundesanstalt für Arbeit. Regensburg, Germany 1971.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.

(Received September 5, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 6, 748–752

SLEEP DEPRIVATION EFFECTS ON THE VESTIBULAR HABITUATION PROCESS[1]

PATRICK J. DOWD[2]

United States Air Force School of Aerospace Medicine,
Brooks Air Force Base, Texas

The effects of sleep deprivation on habituation of the vestibular system in a stressful situation were examined. One hundred forty three experienced pilots were exposed to two tests of Coriolis acceleration after periods of sleep deprivation (24–30 hours) or rest (6 hours). Nystagmus responses to Coriolis stimulation were recorded after four right-to-left tilts. Sleep deprivation resulted in (*a*) increased sensitivity to Coriolis stimulation; (*b*) a decreased recovery rate; and consequently (*c*) interference with the vestibular habituation process. The implications of sleep loss in increasing the hazards of flying are discussed.

Very little information is available on the effects of moderate sleep deprivation on the vestibular sensory system. Yet the interactions between sleep mechanisms and the vestibular system have relevance to a number of areas of human performance involving sensitivity to motion stimuli. Particularly relevant is the problem of sleep loss and the performance of aircraft pilots.

Coriolis acceleration, induced by tilting the head in the frontal plane while rotating about a vertical axis, has been found to produce stress for both experienced pilots and nonflyers (Dowd, 1973). Coriolis is a sensation of rolling, banking, or pitching about a third axis orthogonal to the aircraft's turn axis and the head-movement axis. Prolonged rotary stimulation seems to overdrive the vestibular system resulting in compensatory eye movements, termed nystagmus. The eye movements consist of a slow phase in the opposite direction of rotation with a fast phase in the rotating direction. Vertical nystagmus can be recorded and reduced quantitatively using a simple two-parameter analog function (Cramer, Dowd, Moore, & Collins, 1967).

Nystagmic responses to Coriolis stimuli repeated over days display attenuation of sensitivity, with some habituation. Habituation to a Coriolis stimulus means a reduction of nystagmic responses, subjective sensations, and vestibulo-autonomic reactions to repetitive stimulation. Studies (Dowd, 1973) have indicated that vestibular sensitivity, such as the rate of recovery and the intensity of nystagmus elicited by Coriolis accelerations, is significantly modified by the effects of flying experience. Such modifications are identified as habituation of the vestibular system process.

The purpose of the present study was to determine if moderate sleep deprivation of 24 to 30 hours would modify the pilot's vestibular habituation process by affecting the sensitivity and the recovery rate of the vestibulo-ocular responses to Coriolis acceleration.

METHOD

Subjects

The subjects were 143 experienced Air Force flying personnel undergoing physical evaluations.

Procedure

Two Coriolis test periods were given to all subjects; 48 hours after the initial period all subjects received the repeat test. The subjects were divided into three groups: Group A—44 subjects who were deprived of sleep for the first test and had rest for the repeat test; Group B—38 subjects who had rest for the first test and had no sleep for the repeat test; and Group C—61 subjects who had rest for both tests. All subjects labeled "fatigued" were tested after a sleep deprivation period of 24 to 30 hours. All subjects labeled "rested" had at least 6 hours of sleep prior to testing.

---

[1] The research reported in this article was conducted in the Enivronmental Sciences Division, United States Air Force School of Aerospace Medicine, Aerospace Medical Division, (AFSC), United States Air Force, Brooks Air Force Base, Texas. Further reproduction is authorized to satisfy the needs of the U.S. government. The voluntary informed consent of the subjects used in this research was obtained in accordance with Air Force Regulation 80-33.

The author expresses his appreciation to Dr. B. O. Hartman and Dr. R. L. Cramer for their advice, to Francis A. Brogan and James Bower for their technical assistance, and to Richard McNee for statistical analyses.

[2] Requests for reprints should be sent to Patrick J. Dowd, United States Air Force School of Aerospace Medicine/VNE, Brooks Air Force Base, Texas 78235.

A standardized testing procedure was used (Dowd, 1966). Silver electroencephalogram disc electrodes were mounted above and below each eye, with a ground electrode mounted on the forehead between the eyebrows. Subjects were seated in the stimulator and instructed (*a*) to keep their eyes open, (*b*) to refrain from changing their body or head positions, and (*c*) to refrain from blinking or talking. All stimulations were conducted in a darkened room. Initially, before the beginning of rotation, the subject was tilted to the left, 30 degrees from the vertical. Rotation was always in a counterclockwise direction at 18 revolutions per minute. After 2 minutes of exposure to constant velocity, the subject was tilted through the vertical to 30 degrees to the right. This tilt was completed in 3 seconds. After 2 minutes, the subject was tilted again to his initial position. Alternating tilts continued until a total of eight tilts were completed: four left-to-right and four right-to-left.

Electronystagmograms, recorded during each tilt and for 30 seconds following the cessation of each tilt, were processed and statistically analyzed for each of the four tilts from right-to-left. The right-to-left tilt produces a sensation similar to that perceived by experienced flying personnel during a rolling climb in a high-speed aircraft (Dowd, 1973). The left-to-right tilt stimulus was described by pilots as a dangerous and unpracticed maneuver which they avoided in flight. Visceral disturbances were induced by the left-to-right tilt (Dowd, 1973). Thus, the more neutral stimulus, the right-to-left tilt, was used in the analysis of this study.

Scoring

The magnitude of the nystagmic slow-phase deviations of the eyes in degree of displacement was obtained for each second following cessation of the four stimulus tilts. These were combined into average degrees per second for each subject. The average values were then reduced to fractions of the reference signal and presented as arbitrary units per second.

The exponent of the decay function was obtained by taking the logarithm of the average values of the slow-phase deviation in arbitrary units per second. The slope of the line generated from these data described the rate of recovery (β) from Coriolis stimulation as a nystagmus decayed over time: The steeper the slope, the more rapid the recovery from this disorienting stimulus.

Temporal variations in slow-phase velocity of the vertical nystagmus subsequent to stimulation of the semicircular canals can be described by a negative exponential decay function:

$$R_t = R_o e^{-\beta t},$$

in which t equals seconds after cessation of stimulus, R_t equals the response at time t, and R_o equals response at cessation of stimulus where t equals 0 (maximum eye velocity), β equals a parameter of the system which governs the rate at which activity decays or the slope of the negative exponential function.

This relationship has been demonstrated repeatedly (Dowd, 1973). Inferentially the response during acceleration should develop according to:

$$R_{tg} = \alpha\, S(1 - e^{-\beta\, tg}),$$

where S equals sensitivity of the system, tg equals any instant during acceleration intensity (α). The response (R_o) at the termination of three seconds of acceleration of this tilt and the onset of decay is:

$$R_o = \alpha\, S(1 - e^{-3\beta}).$$

R_o and β can be determined from experimental data on decay. By substituting these values into the last equation, the product of the coefficients of acceleration intensity (α) and the vestibular sensitivity (S) can be determined. For brevity this term (α S) is referred to as the sensitivity product. Also S is defined in terms of the maximum eye velocity (R_o) and the rate of recovery (β), so it is possible, for these experimentally derived values to affect the changes in the sensitivity coefficient (S).

The dynamic characteristic (β) represents the decay of the response or rate of recovery from this stimulus, so the greater the numerical value, the more rapid the rate of recovery. The sensitivity product (α S) describes the system's susceptibility to this stimulus, and its numerical value increases as the sensitivity increases, with acceleration being held constant.

RESULTS

None of the subjects reported or showed motion sickness symptoms. Samples of electronystagmographic tracings of induced vertical nystagmic responses of two subjects, one in Group A (fatigued), the other in Group C (rested), to Coriolis stimulation (right-to-left tilts) are presented in Figure 1. Note that the duration of the nystagmic responses in the initial test of the fatigued subject continued for almost 20 seconds after cessation of the stimulus, and for almost 13 seconds in the repeat test. The rested subject shows a duration of almost 13 seconds for both initial and repeat tests. Compare these frequencies of fatigued test period with the rest test periods.

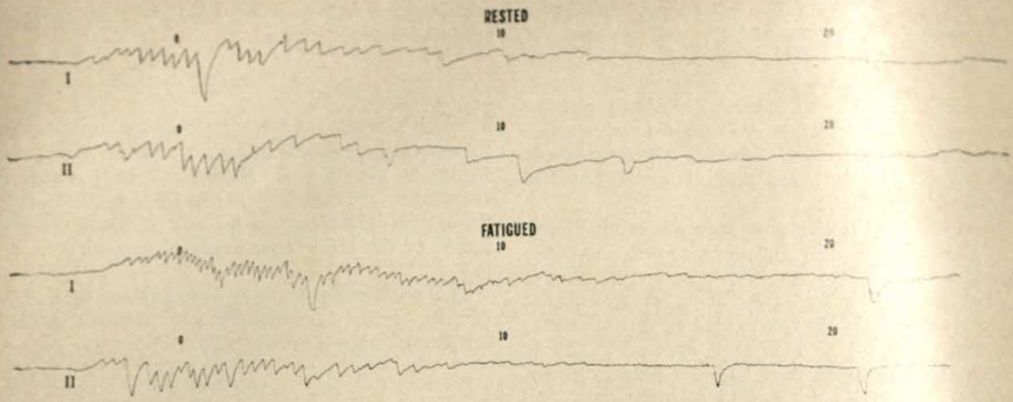


FIGURE 1. Induced vertical nystagmic responses to Coriolis stimulation (right-to-left tilts) of two subjects: one in the rested state, the other in the fatigued state. (I = initial test, II = repeat test, 0 = cessation of stimulus, 10 and 20 = seconds.)

Mean data and standard deviations for the decay function or recovery rate (β) and the sensitivity products (α S) for the initial, repeat, and repeat minus initial test periods for Groups A, B and C appear in Table 1.

The effects of sleep deprivation on the nystagmic responses were estimated by one-half the difference of the differences between the repeat and the initial tests for Groups A and B. These estimates were 2.90 for sensitivity (α S) and −.046 for recovery rate (β). Both of these effects were significant using t tests ($p < .001$).

TABLE 1

MEANS AND STANDARD DEVIATIONS FOR β AND αS

Variable	Group A (44)	Group B (38)	Group C (61)
β (rate of recovery)			
Initial			
$\bar{X}$	.150	.198	.200
SD	.057	.039	.055
Repeat			
$\bar{X}$	.180	.136	.251
SD	.078	.036	.073
Repeat − Initial			
$\bar{D}^{a}$	.030	−.061	.050
SD	.058	.043	.046
α S (sensitivity)			
Initial			
$\bar{X}$	11.19	6.19	6.53
SD	3.71	1.66	2.81
Repeat			
$\bar{X}$	8.68	9.50	5.63
SD	3.42	2.77	2.30
Repeat − Initial			
$\bar{D}$	−2.51	3.30	−.90
SD	3.36	2.39	1.65

[a] $\bar{D} = \bar{X}$ repeat $- \bar{X}$ initial; columns may not compute due to rounding.

To determine the "learning" effect a comparison was made between the repeat and the initial test periods for Group C. A significant decrease in α S and a significant increase in β ($p < .001$) were found.

DISCUSSION

Analysis of the parameters β and α S of the nystagmic measures for both the rested and the sleep-deprived groups indicated that fatigue (moderate sleep deprivation of a period slightly more than 24 hours) exacerbates the vestibular responses to Coriolis stimulation, resulting in interference with the vestibular habituation processes developed by pilots in their flying experience (Dowd, Moore, & Cramer, 1966).

During the repeat rested test, Group A showed an increased recovery rate with less sensitivity in their nystagmic responses. One could speculate that the initial test was novel, and these results were simply due to lack of sleep prior to this test. Group B, however, had knowledge of the test and with lack of sleep prior to the repeat test, their nystagmic responses showed decreased recovery rate with increased sensitivity. These factors resulted in an inhibited vestibular habituation process. Group C showed typical pilot responses with the normal decrease in sensitivity and increase in recovery rate of nystagmic responses to repeat Coriolis stimulation. There was no interference with the vestibular habituation process in Group C's responses.

Past studies (Dowd, 1973; Dowd & Cramer, 1967) have shown that subjects repeatedly exposed to such a Coriolis stimulus displayed a decrement in the sensitivity product and rate of recovery. This vestibulo-cerebellar habituation suppresses nystagmic responses elicited by Coriolis stimulation, and this suppression is enhanced by highly advanced flight experience (Dowd, 1973; Dowd & Cramer, 1967).

In this study the habituation was modified by the stress condition of fatigue or moderate sleep deprivation in that vestibulo-cerebellar sensitivity to such Coriolis accelerations had increased. Similar results were obtained for experienced pilots under another stressful condition, that of Coriolis-induced motion sickness. Pilots susceptible to motion sickness showed an increased sensitivity and a lesser rate of recovery in their nystagmic responses to Coriolis stimulations.

Results of this study are opposite to those expected if minimal arousal from rotation were the only factor affecting sensitivity. Other studies (Collins, 1962; Collins & Guedry, 1962) have clearly demonstrated that mental relaxation (or a state of reverie) strongly suppressed, while alert conditions amplified, vestibular nystagmic responses to rotation; whereas vestibular responses to rotation, while the subjects were sleeping, clearly demonstrated depression of vestibulo-ocular responses. It was inferred that these sleep effects could be "referred to some influence exerted by the sleep mechanism on the brain stem reticular structures intercalated in the vestibulo-ocular reflex pathway [Jones & Sugie, 1972, p. 51]."

It appears that combined stresses from Coriolis acceleration and sleep deprivation are needed to alter the vestibular habituation process of pilots. Scholander (1961) demonstrated supportive findings in that the effects of combined stresses of lack of sleep with "continued pressure for performance" caused the electrodermal response to increase. Scholander (1961) explained this effect was due to a delay of the habituation process by some central neural mechanism that more or less "alerts" the subject to respond to such stress. Scholander's explanation of a "higher level of internalized attention" affecting a central habituation process supports the findings of this report.

It appears that different neural pathways are effectively involved: one by a fatigue condition and the others by either an alert or relaxed state. Reverie does not affect the vestibulo-cerebellar systems' habituation processes. It only does, however, influence the oculomotor system's response to vestibular stimulation such that the individual acts as if distracted and inattentive to the vestibular stimulus. The fatigued state, on the other hand, modifies the neural elements associated in the vestibular system's suppression or habituation process. This, in turn, affects the oculomotor system's response, and the individual has no conscious control over his reactions to vestibular stimuli (Kennedy, 1972; Wolfe, 1968, 1969, 1971).

The role of the cerebellar areas in the regulation of eye movements has been explored by various studies (Cohen, Goto, Shanzer, & Weiss, 1965; Reding & Fernandez, 1968). The results support the hypothesis that specific areas in the cerebellum exert control over the vestibular nuclei and the resulting oculomotor responses. Thus, controlling habituation develops through some direct inhibitory action or facilitory action in neural pathways between these nuclei (Wolfe, 1968).

It is not known whether the processes of fatigue influence orientation per se, or vice versa, or whether they are independent of each other save for the common sensory modalities and neural pathways involved. As reflected by changes in both parameters, recovery rate (β) and sensitivity (α S), fatigue or moderate sleep deprivation induces a potentially hazardous condition for flying. It appears that in a fatigued state, skilled pilots regress to a lesser state of vestibulo-cerebellar habituation. It can be concluded from this study that the vestibular system's habituation process, developed by experienced pilots, actively suppresses the nystagmic responses to such stimulation as Coriolis acceleration. However, this suppression is somehow weakened by a moderate sleep deprivation of a 24- to 30-hour period.

REFERENCES

Cohen, B., Goto, K., Shanzer, G., & Weiss, A. H. Eye movements induced by electrical stimulation of the cerebellum in the alert cat. *Journal of Experimental Neurology*, 1965, **13**, 145–162.

Collins, W. E. Effects of mental sets upon vestibular nystagmus. *Journal of Experimental Psychology*, 1962, **63**, 191–197.

Collins, W. E., & Guedry, F. E., Jr. Arousal effects and nystagmus during prolonged constant angular acceleration. *Acta Otolaryngologica*, 1962, **54**, 349–362.

Cramer, R. L., Dowd, P. J., Moore, E. W., & Collins, F. G. The changing parameters of the habituating vestibular system. United States Air Force School of Aerospace Medicine Technical Report No. 67-85, 1967, Brooks Air Force Base, Texas.

Dowd, P. J. Factors affecting vestibular nystagmus in Coriolis stimulation. *Acta Otolaryngologica*, 1966, **61**, 228–236.

Dowd, P. J. A useful test in selecting motionsick-prone individuals. *Space Life Sciences*, 1973, **4**, 391–401.

Dowd, P. J., & Cramer, R. L. Habituation transference in Coriolis acceleration. *Aerospace Medicine*, 1967, **38**, 1103–1107.

Dowd, P. J., Moore, E. W., & Cramer, R. L. Effects of flying experience on the vestibular system: A comparison between pilots and nonpilots to Coriolis stimulation. *Aerospace Medicine*, 1966, **37**, 45–47.

Jones, G. M., & Sugie, N. Vestibulo-ocular responses in man during sleep. *Electroencephalography and Clinical Neurophysiology*, 1972, **32**, 43–53.

Kennedy, R. S. The relationship between habituation to vestibular stimulation and vigilance: Individual differences and subsidiary problems. NAMRL Monograph #20, July 1972. Naval Aerospace Medical Research Laboratory, Pensacola, Florida.

Reding, G. R., & Fernandez, C. Effects of vestibular stimulation during sleep. *Electroencephalography and Clinical Neurophysiology*, 1968, **24**, 75–79.

Scholander, T. The effects of moderate sleep deprivation on the habituation of autonomic response elements. *Acta Physiologica Scandinavica*, 1961, **51**, 325–342.

Wolfe, J. W. Evidence for control of nystagmic habituation by folium-tuber vermis and fastigial nuclei. *Acta Otolaryngologica, Supplementum 231*, 1968, **231**, 1–48.

Wolfe, J. W. Mesodiencephalic and cerebellar influences on optokinetic and vestibular nystagmus. *Experimental Neurology*, 1969, **25**, 24–34.

Wolfe, J. W. Relationship of cerebellar potentials to saccadic eye movements. *Brain Research*, 1971, **30**, 204–206.

(Received October 15, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 6, 753–758

INTEGRATION OF ATTRIBUTES IN BUS TRANSPORTATION:
TWO MODELING APPROACHES[1]

KENT L. NORMAN[2]
University of Alabama

JORDAN J. LOUVIERE
Department of Geography, Florida State University

Judged probability of riding hypothetical bus systems was studied as a function of fare, frequency of service, and walking distance to the bus stop. The judgment process of 20 paid college students was modeled by two different approaches. A policy capturing approach using regression analysis indicated that a linear model accounted for the data adequately. However, a polynomial model incorporating both linear and multiplicative components in the judgment policy provided a better fit. A method of scaling known as functional measurement served to reject the linear model and confirmed that a multiplying model was correct. Functional measurement provided subjective scales for fare, frequency of service, and proximity to the bus stop. The advantages and limitations of both approaches were discussed.

Administrative decisions based on public preferences have often relied on past experience and guesswork. Planners would like to be able to estimate and predict public response to policy decisions. To do this the planner must know two things. First, he must know which attributes in a proposed system are relevant to public preference and what their relationship is to the perceived desirability of a particular system. Second, the planner needs to know how these different attributes combine or interact in the formation of an overall preference.

Two rather different modeling procedures were considered in this study and applied to a practical question concerning choice of transportation mode. The first method is generally known as "policy capturing" (Hammond, 1955). This method stems directly from Brunswik's (1956) Lens Model and has been popularly applied in a number of areas of psychology. The second method is known as "functional measurement" and derives from Anderson's (1974) work on information integration theory. The similarities and differences between these approaches have been discussed elsewhere in detail (Anderson, 1972; Slovic & Lichtenstein, 1973), but are briefly reviewed in the present context.

The policy-capturing approach usually presents stimulus combinations which cover a "representative" range, and data are analyzed by regression and correlational methods. This means that a physical metric on each attribute is necessary to allow for a correlation with the subject's judgment response. Linear regression models are usually used, though nonlinear or configural models have been recently explored (Goldberg, 1971).

Functional measurement constitutes a scaling theory which allows for a subjective value s for each physical stimulus value and an associated weight w which reflects the importance of that stimulus. Subjective stimulus scales can be derived independently of the physical metric. Although functional measurement has been primarily concerned with linear models such as adding and averaging, it has lately been applied to more complex configural models (Anderson, 1972; Shanteau & Anderson, 1972).

Both policy capturing and functional measurement have been quite successful in modeling judgment behavior, but investigators have enjoyed little communication on these two

---

[1] This research was supported in part by the Institute of Urban and Regional Research at the University of Iowa as part of a program of research and training in Urban Transportation sponsored by the Urban Mass Transportation Administration of the U.S. Department of Transportation. The authors are grateful to Norman H. Anderson for his helpful comments and to Lynn L. Beavers and Frank C. Stetzer for their help in conducting the experiment and analyzing the results.

[2] Requests for reprints should be sent to Kent L. Norman, Department of Psychology, University of Alabama, Box 2968, University, Alabama 35486.

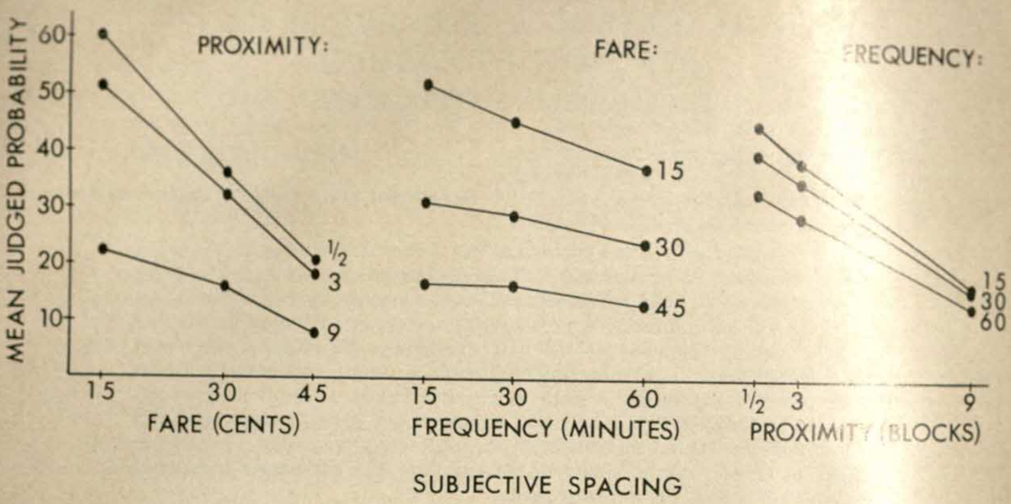


FIGURE 1. Mean judged probability of riding a bus system as a function of fare, frequency of service, and proximity to the bus stop. (Each panel presents a two-way plot by averaging over one factor. Values on the abscissa are subjectively spaced according to a multiplying model.)

approaches. The present study was therefore subjected to both modeling approaches. This comparative technique was intended to provide data on the properties of choice of transportation mode as well as insight into the advantages and disadvantages of each approach in a practical situation.

METHOD

Task

The problem to which these two approaches were applied is one of particular interest to planners of a transportation system. In predicting public response to a new bus system the most desirable experiment would be a direct manipulation of various factors. Historically such experiments have been impractical or impossible. An alternative, especially in the initial stages of a research program, is a paper-and-pencil experiment in which persons rate the probability that they woud ride a bus system described by a set of characteristics.

Hypothetical bus systems varying in fare, proximity to the bus stop, and frequency of bus service were judged in terms of the probability that the individual would ride the described bus system. Subjects were told to make their responses in the context of a trip to or from school. Each attribute describing the bus system could take on any one of three values. Fare was either 15 cents, 30 cents, or 45 cents; proximity to the bus stop was $\frac{1}{2}$ block, 3 blocks, or 9 blocks; and frequency of service was every 15 minutes, 30 minutes, or 60 minutes. All possible combinations of these three attributes were presented as different hypothetical bus systems in a 3^3 factorial design. The 27 combinations were presented in randomized blocks with the order of the three attributes varying nonsystematically. Combinations of attributes were printed in the middle of 5 × 8 inch pages.

Judgments were to be indicated by a slash mark on a 150-millimeter horizontal line on the lower half of each page. Subjects were told that the left end point labeled "0" represented certainty that one would never take a bus with the characteristics given, and that the right end point labeled "100" represented certainty that one would always take the bus described. Intermediate points represented the "proportion of time you would choose to take the bus described." Responses were measured in millimeters and rescaled to range between 0 and 100. End anchors were supplied in the instructions to avoid possible nonlinearity in the response scale caused by responding close to the end points. The end anchors were combinations of attributes more extreme than those actually employed in the experiment. The descriptions composing each randomized block were stapled together in a booklet.

Subjects

Twenty students drawn from the University of Iowa Office of Student Employment served as subjects and were paid $5 each.

Procedure

The experiment was administered in two 1-hour sessions. In a morning session subjects rated bus systems in three booklets at their own rate, rested five minutes, and rated descriptions in an additional three booklets. The same procedure was followed in an afternoon session. The first randomized block of descriptions in each session was treated as practice and was not analyzed, leaving 10 blocks for each subject. Subjects were instructed to treat stimulus sets independently and to proceed directly through the booklet without referring to previous judgments.

RESULTS

Mean judgments are graphed in the panels of Figure 1 as a function of the values of fare and proximity, frequency and fare, and proximity and frequency. It can be seen, as expected, that with increasing fare, judged probability of riding the bus decreased. Likewise, as the time between buses (frequency) and the walking distance to the bus stop (proximity) increased, judgments decreased.

The most striking feature of the data in Figure 1 was the convergence of the lines to the right. This indicated that when one of the factors was low in desirability the other pieces of information had a smaller effect on the judgment. For example in the first panel, if the bus stopped 9 blocks away the probability of riding was only influenced slightly by fare; whereas, fare was quite important in determining the judgment when the bus stopped $\frac{1}{2}$ block away.

Policy-capturing

A linear model of the following form was initially fit to the data:

$$R_{ijk} = b_0 + b_1 S_i + b_2 T_j + b_3 U_k, \quad [1]$$

where R_{ijk} is the predicted response for fare S_i, frequency of service T_j, and proximity to the bus stop U_k, with their respective regres-

TABLE 1

OBSERVED AND PREDICTED RATINGS OF BUS SYSTEMS FOR DIFFERENT MODELS

Stimulus combination			Observed rating	Predicted rating		
				Regression models		Functional measurement multiplying model (R^2 = .997)
Fare (in cents)	Frequency (in minutes)	Proximity (in blocks)		Linear (R^2 = .948)	Polynomial (R^2 = .994)	
15	15	$\frac{1}{2}$	71.01	58.11	67.96	68.56
15	15	3	59.84	50.96	58.26	59.50
15	15	9	23.41	33.80	25.38	26.06
15	30	$\frac{1}{2}$	60.20	55.32	61.64	60.75
15	30	3	52.14	48.17	53.22	52.73
15	30	9	22.71	31.02	23.15	23.10
15	60	$\frac{1}{2}$	49.00	49.74	48.99	50.00
15	60	3	42.29	42.59	43.15	43.39
15	60	9	19.50	25.44	19.25	19.00
30	15	$\frac{1}{2}$	40.72	43.48	43.48	42.05
30	15	3	35.56	36.33	35.78	36.50
30	15	9	17.63	19.18	17.04	16.00
30	30	$\frac{1}{2}$	37.68	40.69	39.54	37.27
30	30	3	32.16	33.54	32.68	32.34
30	30	9	16.12	16.39	15.94	14.17
30	60	$\frac{1}{2}$	29.66	35.11	31.67	30.66
30	60	3	28.05	27.96	26.48	26.62
30	60	9	13.33	10.81	13.76	11.66
45	15	$\frac{1}{2}$	21.81	28.86	21.11	23.30
45	15	3	19.38	21.71	18.81	20.22
45	15	9	8.50	4.56	7.34	8.86
45	30	$\frac{1}{2}$	20.79	26.07	20.91	20.65
45	30	3	19.98	18.92	18.99	17.92
45	30	9	8.43	1.76	8.45	7.85
45	60	$\frac{1}{2}$	18.82	20.49	20.50	16.99
45	60	3	15.14	13.34	19.36	14.74
45	60	9	5.50	−3.82	10.67	6.46

Note. Correlation indexes (R^2) are between observed and predicted ratings.

sion weights b_1, b_2, and b_3, and intercept b_0. The estimates of the regression weights and intercept were −.975, −.186, −2.859, and 76.95, respectively. Table 1 lists the observed and predicted judgments for the linear model. The correlation index (R^2) of .948 might seem more than satisfactory, but a comparison of observed and predicted judgments revealed a systematic underestimation at extreme stimulus values and overestimation at intermediate stimulus values.

The linear regression model did not capture the configural aspect of the data noted in Figure 1 since it assumes that the magnitude of the effect of each factor is independent of the levels of other factors. It was assumed that the addition of cross product terms could increase the fit by incorporating multiplicative components in the judgment policy. Consequently, a third-order polynomial regression equation was applied with all possible cross product terms. The predicted judgments for this model are listed in Table 1 with an R^2 of .994.

The conclusion based solely on regression analysis was that while a linear model accounted for a large proportion of the variance, the subject's judgment policy was best described by a combination of linear and multiplicative components in a polynomial regression model.

Functional Measurement Analysis

Like policy capturing, functional measurement seeks the best fitting model that describes the data. Analysis of variance often helps to indicate the form of the appropriate model. The presence of interactions may indicate a multiplicative relationship between factors (Shanteau & Anderson, 1972). In the present case the three two-way interactions shown in Figure 1 were all highly significant. In addition, the three-way interaction of all factors was also significant ($p < .001$, in each case). The presence of these interactions indicated that a nonlinear model would be necessary to account for these data. The form of the observed interactions suggested a three-factor multiplying model of the following form:

$$R_{ijk} = (s_i^{w_1}) \quad (t_j^{w_2}) \quad (u_k^{w_3}), \qquad [2]$$

where s_i, t_j, and u_k are the estimated subjective values of the levels of fare, frequency of service, and proximity to the bus stop with their respective weights w_1, w_2, and w_3. The weights on these factors are atheoretical, and in the present case are arbitrary due to the interval nature of the subjective scales (Anderson, 1974). To allow a solution for Equation 2, the weights were set equal to 1. Additionally, a common origin for the scales was set such that $s_1 = t_1 = u_1$. Interval scale estimates of the subjective values were found by the method of maximum likelihood (Norman, 1973).

Table 1 lists the predicted judgments for the multiplying integration model. Predicted judgments were very close to the observed values, and the correlation index of .997 indicated that this model fit at least as well as the polynomial regression model. This is interesting to note since the single multiplying model requires the estimation of less than half as many independent parameters.

A further test of the multiplying model was provided since it predicted a fan of straight lines when the factors were subjectively spaced. This is shown in Figure 1 where the abscissas in the three panels are spaced subjectively for fare, proximity to the bus stop, and frequency of service. This graphic result may be statistically tested since the multiplying model predicts that all the interactions should be localized in the bilinear or trilinear components (Shanteau & Anderson, 1972), and residual components should be nonsignificant. This pattern was obtained: for each interaction, the bilinear or trilinear component was significant ($p < .001$, in each case) and the residual was not significant ($p > .20$, in each case).

Figure 2 plots the relationship between the physical scales and the subjective scales for each factor for the averaging and multiplying models. The straight lines represent the assumption of regression analysis that the physical and subjective scales are linearly related. For fare and frequency, the subjective probability of riding the bus was a negative

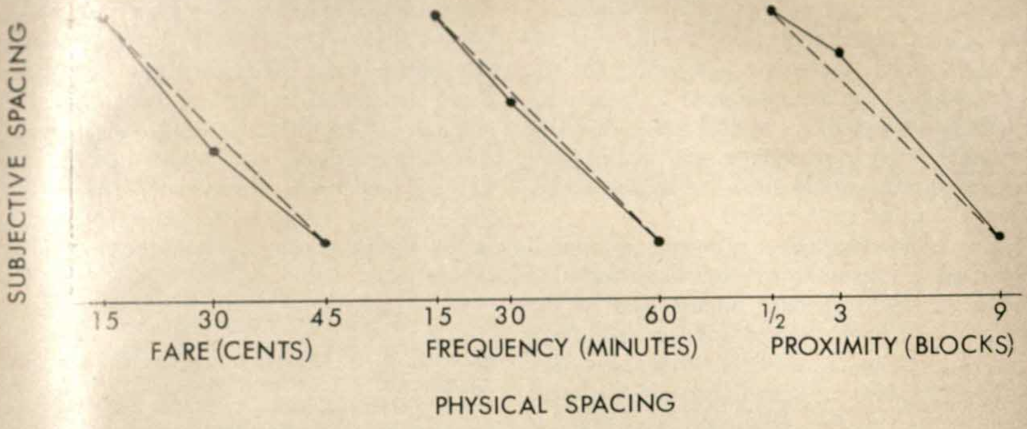


FIGURE 2. Subjective values as a function of the physical metrics of fare, frequency of service, and proximity to the bus stop. (Dashed lines represent a linear relationship between the subjective and physical metrics.)

decelerating function of the physical values, that is, the decrease from 30 cents to 15 cents was subjectively greater than the decrease from 45 cents to 30 cents. A similar relationship held for frequency of service. For proximity to the bus stop, subjective probability was a positively decelerating function so that equal increments along the physical scale lead to increasing decrements in subjective probability.

This analysis based on functional measurement clearly showed that a linear model was inadequate and should be straightforwardly rejected. The multiplying model, on the other hand, provided an excellent fit of the data. Given that the multiplying model is the correct model, functional measurement guarantees that the estimated subjective values (see Figure 2) constitute valid interval scales for fare, frequency of service, and proximity to the bus stop.

DISCUSSION

Planners are interested in the two questions posed earlier regarding the effects of system attributes and the integration rule. First, it is clear that the three attributes of fare, frequency of service, and proximity to the bus stop had significant effects on rating judgments. No statement about the relative importance of the three attributes can be made independent of the levels of the attributes chosen in this experiment. However, within the levels chosen, fare had the greatest effect and frequency of service the least. Also, it appeared that these scales were not necessarily linear with subjective ratings. One would exepct that as distance to the bus stop increased to prohibitive levels, the subjective ratings would no longer change appreciably since they would be close to zero.

Second, the integration model is multiplicative in form. If any attribute is at an extremely low value the overall rating will be low despite the values of other attributes, but it is also compensatory since a high-valued attribute may compensate for a moderately low value on another attribute. Given such a model as a description of the integration process, planners may take the next step of optimizing bus ridership at minimum cost levels. One needs only to relate the subjective values to the cost of the system and find those levels that maximize system desirability at a predetermined cost. As always, caution must be exercised in generalizing the results of a contrived experiment to a real situation.

Evaluation of Approaches

Regression analysis and functional measurement differ in several fundamental assumptions. Regression analysis must assume a particular relationship between the physical values and the subjective values, since physical values are used in the prediction equation. Although physical metrics were available in the present experiment, there are many situations in which there is no objective metric. Furthermore, the assumption of a linear relationship between the physical and subjective values is often unreasonable and in the present experiment would have led to the acceptance of an incorrect model.

The policy-capturing approach produced confusing results as to whether the integration model of the subjective impressions in the present experiment was additive or multiplicative. A polynomial model provided the best fit by incorporating both linear and nonlinear components in the judgment. However, it should be mentioned that the generality of a high-order polynomial may be limited. The coefficients of higher order terms which increased predictability in the present situation may be unreliable and may decrease predictability if applied in other situations. Furthermore, the psychological interpretation of the polynomial model is lacking. It is not at all clear what processes or combinations of processes the higher order coefficients reflect.

Nevertheless, the policy-capturing approach has permitted high levels of predictability (Slovic & Lichtenstein, 1973) and has been particularly useful when the data are correlational in nature and when a factorial manipulation of the variables is not feasible. However, when it comes to diagnosing the correct integration model, the policy-capturing approach may not fare so well.

In contrast, functional measurement does not require a physical metric and makes no assumption concerning the relationship between the physical and subjective metric. Unlike regression analysis, functional measurement constitutes a theory of measurement which can guarantee valid interval scales of the subjective stimuli (Anderson, 1974). As seen in Figure 2, the spacing between subjective values is not the same as the spacing between the respective physical values. Reliance on analysis of variance served to reject the incorrect linear model and indicated that a multiplying model was appropriate. The conclusion based on functional measurement is clear and straightforward—the data are multiplicative in form.

REFERENCES

Anderson, N. H. Looking for configurality in clinical judgments. *Psychological Bulletin,* 1972, **78,** 93–102.

Anderson, N. H. Algebraic models in perception. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception,* Vol. 2. New York: Academic Press, 1974.

Brunswik, E. *Perception and the representative design of experiments.* Berkeley: University of California Press, 1956.

Goldberg, L. R. Five models of clinical judgment: An empirical comparison between linear and nonlinear representations of the human inference process. *Organizational Behavior and Human Performance,* 1971, **6,** 458–479.

Hammond, K. R. New directions in research on conflict resolution. *Journal of Social Issues,* 1965, **21,** 44–66.

Norman, K. L. *A method of maximum likelihood estimation for information integration models.* (CHIP Report 37) San Diego, Calif.: Center for Human Information Processing, August 1973.

Shanteau, J., & Anderson, N. H. Integration theory applied to judgments of the value of information. *Journal of Experimental Psychology,* 1972, **92,** 266–275.

Slovic, P., & Lichtenstein, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. In L. Rappoport & D. A. Summers (Eds.), *Human judgment and social interaction.* New York: Holt, Rinehart and Winston, 1973.

(Received November 12, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 6, 759–760

RELATIONSHIP BETWEEN CONSUMERS' CATEGORY WIDTH AND TRIAL OF NEW PRODUCTS:

A REAPPRAISAL

LYMAN E. OSTLUND[1]
Graduate School of Business, Columbia University

BRYNULF TELLEFSEN
Department of Administrative Science and Economics, State University of New York-Plattsburgh

An earlier study conducted in 1973 by Donnelly, Etzel, and Roeth concluded "an individual's breadth of categorization is related to his purchase of new products [p. 337]." Questions are raised about how the products studied were selected, choice of subjects, and the measure of association used. It is argued that the original conclusion is weakly supported.

In a recent study, Donnelly, Etzel, and Roeth (1973) concluded that a consumer's "breadth of categorization" is related to his purchase of new products. It is our belief that this conclusion is not justified by their study, due to a number of methodological problems.

Procedure

One problem with the study is the definition of *new products*, which is based on four product characteristics or dimensions: packaging, physical appearance, required user preparation, and manufacturer's technological processing. While the four-factor model of product newness may ultimately prove useful, the reader is not told how each factor was measured and the direction of relationship of each factor to *newness*. While a radical departure in packaging approach would undoubtedly be perceived by most housewives as representing newness, how is physical appearance supposed to operate? Is "required user preparation" supposed to be measured in time, effort or what? Must a product offer less required user preparation to be considered new? Does more "manufacturer's technological processing" mean more newness, or is it more sophisticated processing that means more newness? Can the researchers accurately judge manufacturer processing in such terms? Even if she can, is it likely that a housewife actually attempts to assess manufacturer's technological processing in identifying and/or evaluating new products? Finally, Donnelly et al. (1973) give no account of how the factors are combined (multiplied? added?) in assessing newness, and whether or not a product must score a certain minimum on each factor to be considered new. In short, we need to know more about how their theory of newness was applied before the selection of their five food product groups for study can be evaluated.

The five product groups selected for study were (*a*) frozen pudding, (*b*) frozen donuts, (*c*) canned pudding, (*d*) canned cake frosting, and (*e*) instant oatmeal with fruit. It would appear that four of these products represent desserts, and are new only in the sense of how they are preserved, that is, frozen or canned. Thus, according to the scheme developed by Robertson (1971), these products would appear to involve "the least disrupting influence on established consumption patterns. Alteration of a product is . . . involved, rather than the creation of a new product [p. 7]."

For these reasons, we believe the Donnelly et al. (1973) study cannot claim to have necessarily identified five product groups of significant newness as perceived by housewives.

Subjects

In their sampling of housewives, the researchers screened for familiarity with the five product groups. If a housewife "exhibited a lack of familiarity with some or all of the products [p. 336]," she was dropped from the sample. Since only 17% acknowledged a lack of familiarity with one or more of the five product groups, it is at least possible that a yea-saying bias exists in the dependent variable, namely, *recalled* purchase. Actual purchase data were not available. In this connection the hypothesis is incorrectly stated as "willingness to buy" when in fact the measure is one of recalled behavior.

---

[1] Requests for reprints should be sent to Lyman E. Ostlund, who is now at the Marketing Department, the Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19174.

By having screened out the 35 subjects who disclaimed familiarity with one or more of the product groups, Donnelly, et al. (1973) lost an opportunity to enrich the conceptual utility of their study. Namely, "width of categorization", as a measure of cognitive style, may have at least as much to do with determining familiarity or nonfamiliarity with a product group as it may have to do with determining purchase and nonpurchase within a product group. It would have been useful to relate the width of categorization measure for each respondent to familiarity/nonfamiliarity with each product group before relating it to recalled purchase incidence.

Analysis

Even if we assume that the criteria of product newness and the sampling were adequate, the authors should have used for their analysis a statistical test of association, given their hypothesis, rather than chi-square, which is a test of independence. The lambda test would have been most appropriate, given that the dependent variable is scaled nominally and causality between width of categorization and recalled purchase is implicit in the hypothesis. However, for frozen pudding, frozen donuts, and instant oatmeal with fruit, the lambda coefficient would equal zero; for canned pudding and for canned cake frosting, lambda would equal .02 and .11, respectively.

A less demanding, nondirectional measure of association would be correlation analysis calculated from the chi-square statistic. Converting the Donnelly et al. (1973) chi-square statistics into correlations, we find that they range from .10 to .28 across the five products, indicating a maximum of 8% of the variance in recalled purchase behavior accounted for by width of categorization. Thus, using the authors' own data, there is, at best, only weak evidence that category width predicts purchase of new products.

REFERENCES

Donnelly, J. H., M. J. Etzel, & S. Roeder. The relationship between consumers' category width and trial of new products. *Journal of Applied Psychology*, 1973, **57**, 335–338.

Robertson, T. S. *Innovative behavior and communication*. New York: Holt, Rinehart & Winston, 1971.

(Received October 26, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 6, 761–763

ATTEMPTING TO OPERATIONALIZE PRODUCT NEWNESS:

A REPLY TO OSTLUND AND TELLEFSEN[1]

JAMES H. DONNELLY, JR.,[2] AND MICHAEL J. ETZEL

Deparment of Business Administration, College of Business and Economics,
University of Kentucky

Ostlund and Tellefsen raised questions regarding some methodological procedures, including the operationalization of product newness, employed in an earlier study by Donnelly, Etzel and Roeth. These questions are addressed in an attempt to clarify the procedures and conclusions of the Donnelly et al. investigation.

In their (1974) study Ostlund and Tellefsen raise several points in connection with our study of new product trial (Donnelly, Etzel, & Roeth, 1973). Their principal argument is with our attempt to operationalize product newness. Specifically, they question: (*a*) the definition of "newness" used in the study and (*b*) the choice of products included in the study. In addition, they challenge the use of recall as a measure of purchase and the choice of a statistical test. Each of these points deserves a reply.

Ostlund and Tellefsen (1974), using Robertson's (1971) well-known continuum of new products, have concluded that the products in our study are at best only slightly new (continuous innovations). Though Robertson's work has contributed significantly to thinking and research on new products, his three categories are operationally vague. For example, if consumption is viewed as *purchase*, the housewife buying canned cake frosting may view it as a very new product (a discontinuous innovation). However, if consumption is viewed as the family eating the frosted cake, it is likely that the family members would see little difference between this and other frosted cake eaten previously, and the product becomes at best a continuous innovation.

A failure to define objective criteria to operationalize product newness is a common weakness in many new product studies (Lambert, 1972; Ostlund, 1969; Ostlund, 1972a; Peterson, 1973). It would seem that an attempt to operationalize newness should be embraced and, if possible, improved by researchers interested in new product research.

To overcome problems associated with the selection of products in new product studies, we used four dimensions—packaging, physical appearance, required user preparation, and manufacturer's technological processing—to operationalize product newness. These four dimensions were selected for two reasons. First, they are nonsubjective in nature in that they are tied to the physical characteristics of the products and could therefore be easily determined. Second, they are attributes common to all grocery products and are likely to be promoted by manufacturers and noticed by consumers.

Admittedly, a stronger research design would have incorporated the respondents' evaluation of the degree of newness of the products. However, this was not done for two reasons. First, having respondents report purchase behavior and evaluate newness could produce a bias. For example, respondents may have had a tendency to rate products they had purchased as newer than products not purchased. Second, in order to have several products generally viewed as new, many products would have had to have been included on the questionnaire. Given the length and difficulty of the Pettigrew (1958) scale, it was felt that the addition of a lengthy list of product evaluation statements would severely lessen respondent cooperation.

---

[1] This article is published out of turn at the request of the editor.

[2] Requests for reprints should be sent to James H. Donnelly, Jr., Department of Business Administration, College of Business and Economics, University of Kentucky, Lexington, Kentucky 40506.

Ostlund and Tellefsen (1974) ask, "how each factor (dimension) was measured and the direction of the relationship of each factor (dimension) to newness [p. 759]." Numerous products were compared to the nearest previously available substitutes and were classified as being either different or not different on each dimension. The products included in the study were judged to be significantly different on all four dimensions.

It is surprising to us that Ostlund and Tellefsen (1974) could find a difference on only one dimension (how the products are preserved, or in our terminology, packaging) for the five products in the study. Our observations are quite different. To avoid redundancy, we cite only one example.

The nearest available substitute for canned cake frosting is a powdered frosting mix packaged in a box. Thus, on our four dimensions the two products differ in packaging (can versus box), physical appearance (powder versus semiliquid), user preparation (opening the can and spreading versus adding ingredients and mixing), and manufacturer's technological processing (a semiliquid requires different processing than a powder). Though we feel Ostlund and Tellefsen (1974) were rather conservative in their attempt to match the products and dimensions, our disagreement points to a more important issue. Studies of new products should, whenever possible, incorporate consumers' perceptions of newness rather than researchers'. Possibly future new product research should incorporate matched samples, one to develop a set of products and the other to test the research hypotheses.

Ostlund and Tellefsen (1974) are concerned about the direction of change: "Must a product offer less required use preparation to be considered new [p. 759]" and "does more 'manufacturer's technological processing' mean more newness [p. 759]." We contend that the direction of change from existing products is irrelevant. Webster's New World Dictionary (1961) defines *newness* as "different from one in the past." We concur with this definition. This is consistent with many product changes which may or may not be product improvements but which manufacturers advertise as "new." The actual consumer benefit derived from the change may be questionable; however, the purpose is to create the impression of newness in consumers' minds.

Ostlund and Tellefsen (1974) ask, "How is physical appearance supposed to operate [p. 759]." Physical appearance is simply how the product looks to consumers. Though no other change may be apparent, changing the physical appearance of the product can create the impression of newness. For example, instant coffee crystals differ from powdered instant coffee only in appearance. In our study, all of the products have a physical appearance different from that of previously available substitutes.

Ostlund and Tellefsen (1974) question whether "a housewife actually attempts to assess 'manufacturer's technological processing' . . . (p. 759)." Certainly this remains to be empirically determined. However, we feel that though consumers may not use this terminology, they are frequently informed through advertising about new formulations, processing methods, and technological advances. Thus, technological change is a relevant dimension of newness.

In our study purchase recall was used as a surrogate measure of actual purchase. Ostlund and Tellefsen (1974) point out that this may have produced a "yea-saying" bias. We acknowledge this possibility. However, like other new product researchers who have also used recall (Darden & Reynolds, 1974; Lambert, 1972; Ostlund, 1972b) and other purchase surrogates such as purchase intention (Ostlund, 1972a), we found that practicality made such a measure necessary. We also point to the exploratory purpose of our study as a further justification for a less than perfect measure.

In regard to the analysis of the data, Ostlund and Tellefsen (1974) contend that a measure of association would have been more appropriate than chi-square, a measure of independence. Our results, using chi-square, establish rather conclusively that the relationship between breadth of categorization and purchase of selected new products is not the result of chance. Using correlation analysis, Ostlund and Tellefsen have shown that breadth of categorization explains only a small portion of the variance in recalled purchase. This is not a particularly surprising finding, since we did not in the article and do not now contend that breadth of categorization completely explains new product purchase behavior. We only offer it as one of the many pieces to be considered in the puzzle of explaining who purchases new products.

REFERENCES

Darden, W. R., & F. D. Reynolds. Backward profiling of male innovators. *Journal of Marketing Research*, 1974, 11, 79–85.

Donnelly, J. H., Etzel, M. J., & Roeth, S. The relationship between consumers' category width and trial of new products. *Journal of Applied Psychology*, 1973, 57, 335–338.

Lambert, Z. V. Perceptual patterns, information handling, and innovativeness. *Journal of Marketing Research*, 1972, 9, 427–431.

Ostlund, L. E. The role of product perceptions in innovative behavior. In P. R. McDonald (Ed.), *Marketing involvement in society and the economy*. Chicago, Ill.: American Marketing Association, 1969.

Ostlund, L. E. A study of innovativeness overlap. *Journal of Marketing Research*, 1972, 9, 341–343. (a)

Ostlund, L. E. Are there generalized innovators? *Proceedings of the Midwest Conference, American Institute for Decision Sciences*, 1972, 21–24. (b)

Ostlund, L. E., & Tellefsen, B. The relationship between consumers' category width and trial of new products: A reappraisal. *Journal of Applied Psychology*, 1974, 59, 759–760.

Peterson, R. A. A note on optimal adopter category determination. *Journal of Marketing Research*, 1973, 10, 325–329.

Pettigrew, T. The measurement and correlates of category width as a cognitive variable. *Journal of Personality*, 1958, 26, 532–544.

Robertson, T. S. *Innovative behavior and communication*. New York: Holt, Rinehart & Winston, 1971.

Webster's New World Dictionary. New York: World Publishing Company, 1961.

(Received August 2, 1974)

Manuscripts Accepted for Publication in the *Journal of Applied Psychology*

Consumer Installment Credit Risk: A Need for Criterion Refinement and Validation. Paul M. Muchinsky (Department of Psychology, Iowa State University, Ames, Iowa 50010).

Relationship of Internal–External Control to Work Motivation and Performance in an Expectancy Model. Laurie A. Broedling (Navy Personnel Research and Development Center, Department of the Navy, San Diego, California 92152).

Comparative Analysis of Expectancy and Heuristic Models of Decision Behavior. John E. Sheridan (School of Business Administration, Wayne State University, Detroit, Michigan 48202), Max D. Richards, and John W. Slocum, Jr.

On Simultaneous Monitoring and Tracking. Earl L. Wiener (University of Miami, 222 Engineering Building, Coral Gables, Florida 33124).

Personality and Situational Determinants of Leader Behavior. Robert W. Rice (Department of Psychology, University of Utah, Salt Lake City, Utah 84112) and Martin M. Chemers.

Relative Importance of Applicant Sex, Attractiveness, and Scholastic Standing in the Evaluation of Job Applicant Résumés. Robert L. Dipboye, Howard L. Fromkin (Krannert Graduate School of Industrial Administration, Purdue University, West Lafayette, Indiana 47907), and Kent Wiback.

Inventoried Vocational Interests of Cartographers. Arthur L. Benton (Health and Human Services, Bangor Community College, Caribou Hall, University of Maine, Bangor, Maine 04401).

Cola and Diet Cola Identification and Level of Cola Consumption. Sam H. Lane (Department of Psychology, Cleveland State University, Cleveland, Ohio 44115), James Zychowski, and Kenneth Lelic.

Ecosystem Distrust and the Hard-to-Employ. Harry C. Triandis (Department of Psychology, University of Illinois, Champaign, Illinois), Jack M. Feldman, David E. Weldon, and William M. Harvey.

Job Characteristics and Job Attitudes: A Multivariate Study. Eugene F. Stone (School of Economics and Management, Oakland University, Rochester, Michigan 48063) and Lyman W. Porter.

Journal of Applied Psychology
1974, Vol. 59, No. 6, 764–766

SHORT NOTES

USING SUBORDINATE RATINGS TO ELICIT BEHAVIORAL CHANGES IN SUPERVISORS[1]

W. HARVEY HEGARTY[2]
Indiana University

The purpose of this study was to determine whether feedback of subordinates' ratings of supervisors leads to positive changes in the supervisors' behavior. The employees of the 56 supervisors in the experimental and control groups completed an information-opinion survey using a modified Likert scale format. Feedback reports were given to the experimental supervisors and a second survey was conducted 10 weeks later to measure change. Two independent measures showed significant positive change for the experimental supervisors.

The purpose of this study was to determine whether feedback of subordinates' ratings to their supervisor leads to subordinates perceiving positive changes in their supervisor's subsequent behavior.

From a learning standpoint, the feedback report can provide the supervisor with knowledge of how well he is performing as a supervisor. He may change because he is motivated to become a better supervisor and/or the feedback report reveals some shortcomings that are in contrast to his concept of a good supervisor.

Blum and Naylor (1968) provide a good summary of the types of feedback based on the kind and source of information involved. This study was concerned with extrinsic feedback in which the cues are external to the organism. Since the information about the performance was from a source outside the task, and since there was a time delay between behavior and feedback, this feedback was *secondary* and *summary* in nature. Behavioral statements directly relating to supervisory performance were used, so feedback tended to be more *specific* than *general*.

The earliest study concerning feedback of subordinates' ratings to their supervisors was one conducted by Esso Research in the late 1950s (Maloney & Hinrichs, 1959). Esso supplied its technical supervisors with personal analyses made by their subordinates. Another study in a midwestern power company used responses gathered on an attitude survey of subordinates to feedback to supervisors (Baumgartel, 1959). Wakely (1964) used the feedback method in connection with an actual–predicted ratings approach. These studies all reported favorable results, but did not employ control groups in their research designs.

Two studies reported by Gage have dealt with using feedback to elicit behavior change in an academic setting. On a retake of the teacher ratings, 10 of the 12 items used in this study showed changes in the hypothesized (toward the ideal) direction. A similar study has also been carried out using teacher ratings of their own principal and an ideal principal (Daw & Gage, 1967; Gage, 1963).

METHOD

The subjects for this study were 56 supervisors of staff employees at the University of North Carolina. Thirteen departments at the university were involved in the study. The subjects were the first-level supervisors in each of these departments. No supervisor with fewer than five subordinates was included, and the average number of subordinates was nine.

The 56 supervisors were randomly assigned to the experimental group and the control group. The control group received no feedback between the first and second survey. The experimental subjects received feedback on their behavior two weeks after the first survey. The control supervisors did not expect immediate feedback; the experimental supervisors knew they would receive feedback shortly after the first survey. No subordinate knew who would be receiving feedback and who would not.

The specific behaviors to be measured were 15 4-point Likert scale statements taken from a larger attitude questionnaire. The statements under consideration in this study were all concerned with supervisory behavior. A summary item measuring overall performance on a 7-point scale ranging from poor to excellent was also used.

Complete anonymity was assured by asking each individual to identify himself by using a code that specified only his department and work group. Participation in the survey was encouraged, but employees were told if they had reservations about participating for any reason, they could elect not to participate. The participation rate was 93%.

---

[1] The data for this study were collected in conjunction with the author's doctoral dissertation at the University of North Carolina at Chapel Hill.

[2] Requests for reprints should be sent to W. Harvey Hegarty, Graduate School of Business, Indiana University, Bloomington, Indiana 47401.

TABLE 1

POSTTEST MEANS OF SUBORDINATES' RATINGS OF SUPERVISORS IN CONTROL GROUP AND EXPERIMENTAL GROUP

Item	Control group ($n = 28$)	Feedback group ($n = 28$)
1. I am kept advised by my supervisor as to how I am performing my job.	1.76[a]	2.00
2. I have confidence in the ability of my immediate supervisor.	1.88	2.20
3. The good work I do is often overlooked, but my mistakes are highlighted.	1.38	1.49
4. My supervisor gives clear explanations and instructions.	1.76	1.93
5. I feel free to talk to my supervisor about my problems, ideas, and suggestions.	2.10	2.24
6. My supervisor does not give credit where credit is due.	1.51	1.71
7. I have a clear understanding of what my supervisor expects me to achieve on my job.	1.96	2.05
8. My supervisor "breathes down our necks" too much.	1.92	2.12
9. My supervisor brings out the best in people.	1.57	1.78
10. My supervisor is not considerate of me as a person.	1.95	2.10
11. My supervisor handles things which he should delegate to others.	1.49	1.68
12. My supervisor explains the reasons for rejection or acceptance of my ideas and suggestions.	1.73	1.85
13. My immediate supervisor is quick to take care of complaints.	1.61	1.86
14. My immediate supervisor is interested in me as a person.	1.70	1.86
15. My supervisor consults us when making a decision that affects us.	1.91	2.05

[a] Negatively worded item scores have been reversed so that across all items a higher score is more desirable.

TABLE 2

POSTTEST MEANS OF SUMMARY ITEM

Item	Control group ($n = 28$)	Feedback group ($n = 28$)
Overall, I think my supervisor's performance is excellent . . . good . . . fair . . . poor.	3.45[a]	4.02

[a] Higher score reflects a more positive rating; for both groups $p < .02$.

The experimenter went through a standardized procedure in explaining the results of the feedback to the supervisors when they received their feedback reports. Approximately 60 to 90 minutes was spent with each supervisor in going over his individual report.

Ten weeks after the supervisors were given the feedback reports (12 weeks after the first survey), their employees were asked to participate in a second survey. The format of the second survey was the same as the first.

RESULTS AND DISCUSSION

Table 1 shows the results of comparing the control group with the feedback group. The sign test was used on the posttest scores of the 15 behavioral statements. The experimental group had a higher rating on all 15 items ($p < .001$). The sign test was also used to pretest scores between the two groups; there was not a significant difference at this time.

A t test was used to determine if the experimental group was superior to the control group on the summary item (see Table 2). The experimental group was superior ($p < .02$). There was no significant difference on this item or pretest scores.

The major finding as a consequence of this study was that giving supervisors feedback reports improves the supervisors' performance in the eyes of their subordinates. This was supported by two independent measures of performance as noted above.

It is significant to note that there was a positive change across all 15 behavioral items. The lack of independence among these items precluded a test of significance for each item. However, when a t test is applied, the six items with the greatest amount of change had a $p < .10$ or better.

The focus of the study reported here is on the subordinates' perceptions of their supervisor's behavior. A previous study (Hegarty, 1973) of their own change showed that supervisors also thought that they had become better supervisors as a result of the feedback process.

In summation, this project offers strong evidence that positive change does occur when supervisors find out how their subordinates view their behavior. Some questions that might be addressed in the future along these lines include:

1. How often should the feedback process be used to keep supervisory performance high over a longer period of time?
2. What gains can be made in better supervisory performance if a training program accompanies the feedback process?
3. Are there other measures of performance that could be used to provide feedback (e.g., peer ratings, ratings by the supervisor's superior, and work group productivity)?
4. What are the comparative effects of feedback from different sources?

This researcher feels that these and other relevant questions should be intensively studied because of the potential value of the feedback process to an organization. Maximizing the potential of feedback rests with discovering the most judicious way to incorporate this tool in the continuing growth and development of the supervisor.

REFERENCES

Baumgartel, H. Using employee questionnaire results for improving organizations. *Kansas Business Review,* 1959, **12**, 2–7.

Blum, M. L., & Naylor, J. C. *Industrial psychology: Its theoretical and social foundations.* New York: Harper & Row, 1968.

Daw, R. W., & Gage, N. L. Effect of feedback from teachers to principals. *Journal of Educational Psychology,* 1967, **58**, 181–188.

Gage, N. L. A method for improving teacher behavior. *Journal of Teacher Education,* 1963, **14**, 261–266.

Hegarty, W. H. Supervisor's reactions to subordinates' appraisals. *Personnel,* 1973, **50**, 32–35.

Maloney, P. W., & Hinrichs, J. R. A new tool for supervisory self-development. *Personnel,* 1959, **36**, 46–53.

Wakely, J. H. One way to get meaningful results from attitude surveys. *Personnel,* 1964, **41**, 43–58.

(Received February 15, 1974)

ANNOUNCEMENT

As of June 1, 1974 the editorial office of the *Journal of Applied Psychology* moved to a new address. Please send all manuscripts and correspondence regarding manuscripts to Dr. Edwin A. Fleishman, Editor, American Institutes for Research, 3301 New Mexico Avenue, N.W., Washington, D.C. 20016.

Journal of Applied Psychology
1974, Vol. 59, No. 6, 767–770

CROSS-SAMPLE STABILITY OF THE SUPERVISORY BEHAVIOR DESCRIPTION QUESTIONNAIRE

ANDREW D. SZILAGYI
University of Houston

HENRY P. SIMS, Jr.[1]
Indiana University

A factor analysis was performed on the item intercorrelations obtained from the administration of Fleishman's Supervisory Behavior Description (SBD) to a predominately female sample in a nonprofit organization. Factor congruence coefficients were calculated in order to compare the results from this sample with previous leader behavior studies. The SBD measures of consideration and initiating structure were found to be stable across time, sex differences, type of organization, culture, and analysis technique.

The Supervisory Behavior Description (SBD) questionnaire, first described by Fleishman (1951, 1953, 1957, 1972) has received widespread application as a measure of the well-known leader behavior dimensions of Initiating Structure and Consideration (Stogdill & Coons, 1957). The original form of the SBD was based on a factor analysis of leader behavior description items administered to industrial employees who described their supervisors. Tscheulin (1973) recently reported a replication of Fleishman's factor analysis of the SBD items for a sample of industrial supervisors and managers in Germany. He concluded that the two main factors of Consideration and Initiating Structure were confirmed as relevant dimensions for describing the leader behavior in organizations in West Germany.

The purpose of this present study is to report a replication of Fleishman's factor analysis for a predominantly female sample in a nonprofit environment. The results are compared with those presented by Fleishman and Tscheulin.

METHOD

Fleishman's SBD questionnaire was administered to 1,161 employees of a major medical center located in the Midwest, who used the questionnaire to describe their own supervisors. Administration of the questionnaire was conducted during regular working hours at the center. The sample consisted of professional nurses, therapists, technologists, technicians, secretaries, food and service workers, and administrative employees. Of the respondents 79.6% were female employees. For each item in the SBD, a 5-point Likert scale response categorization (e.g., always, often, occasionally, seldom, never) is used to indicate the frequency with which the supervisor engages in that particular behavior.

RESULTS

The responses to the 1,161 questionnaires were subjected to a principal-axis factor analysis, using the subprogram FACTOR from the *Statistical Package for the Social Sciences* (*SPSS*) (Nie, Bent, & Hull, 1970). The method of factoring was PA 2, principal factoring with iteration, with orthogonal rotation using the varimax method.[2] The factor loadings for the two-factor solution are presented as Sample C in Tables 1 and 2. Also presented in Tables 1 and 2 are the original factor loadings obtained by Fleishman (Sample A) and the recent study by Tscheulin (Sample B). Split-half reliabilities, corrected by the Spearman–Brown formula, were .93 for Consideration, and .89 for Initiating Structure. The two factors explained 39.6% of the variance.

The results indicated a high degree of correspondence between each of the factors in the three studies. In order to quantify the similarities of the results of each of the studies, factor congruency coefficients (phi-coefficients) were computed, using the Wrigley and Neuhaus procedure described in Harman (1970, pp. 256–259). The results, reported in Table 3, reveal phi-coefficients above .90 for the Consideration scale and above .80 for the Initiating Structure scale. The correspondence between the German sample

---

[1] Requests for reprints should be sent to Henry P. Sims, Jr., Room 630–Business, Indiana University, Bloomington, Indiana 47401.

[2] The original method of factor analysis used by Fleishman (1953) was the Wherry–Winer iterative procedure (1953); Tscheulin (1973) used the orthogonal varimax rotation method.

TABLE 1

COMPARISON OF SUPERVISORY BEHAVIOR DESCRIPTION QUESTIONNAIRE ITEM FACTOR LOADINGS ACROSS SAMPLES

Consideration items	Consideration (Factor I)			Structure (Factor II)		
	Sample[a]					
	A	B	C	A	B	C
1. He/she refuses to give in when people disagree with him/her.	−.68	−.38	−.47	.06	.23	.22
4. He/she does personal favors for the employees under him/her.	.40	.65	.27	.06	[illegible]	.36
5. He/she expresses appreciation when one of us does a good job.	.70	.55	.72	.19	[illegible]	.08
7. He/she is easy to understand.	.70	.63	.73	.13	[illegible]	−.02
9. He/she demands more than we can do.	−.40	−.45	−.58	−.08	[illegible]	.42
10. He/she helps his/her people with their personal problems.	.32	.54	.54	.05	[illegible]	.22
11. He/she criticizes his/her people in front of others.	−.49	−.60	−.61	.03	[illegible]	.32
15. He/she stands up for his/her people even though it makes him/her unpopular.	.54	.71	.71	.08	[illegible]	.18
17. He/she insists that everything be done his/her way.	−.52	−.38	−.60	−.01	[illegible]	.45
18. He/she sees that an employee is rewarded for a job well done.	.70	.40	.63	.05	[illegible]	.14
19. He/she rejects suggestions for changes.	−.62	−.54	−.65	−.06	.24	.27
21. He/she changes the duties of people under him without first talking it over with them.	−.69	−.56	−.64	.09	.25	.21
22. He/she treats people under him/her without considering their feelings.	−.72	−.70	−.80	.41	.34	.24
24. He/she tries to keep the employees under him in good standing with those in higher authority.	.68	.71	.49	.17	.30	.25
25. He/she resists changes in ways of doing things.	−.57	—	−.61	−.16	—	.29
26. He/she "rides" the employee who makes a mistake.	−.61	−.51	−.67	.37	.43	.41
28. He/she refuses to explain his/her actions.	−.72	−.45	−.73	.23	.19	.21
29. He/she acts without consulting his/her people first.	−.73	−.67	.17	.01	−.01	.17
31. He/she stresses the importance of high morale among those under him/her.	.73	.34	.38	−.11	.54	.38
32. He/she backs up his/her people in their actions.	.62	.64	.16	.16	.18	.16
33. He/she is slow to accept new ideas.	−.66	−.21	.15	−.06	.04	.15
35. He/she treats all his/her people as his/her equal.	.66	.69	−.12	.28	−.08	−.12
36. He/she criticizes a specific act rather than a particular individual.	.63	.19	.18	.14	−.07	.18
39. He/she is willing to make changes.	.78	.55	.05	.09	−.10	.05
40. He/she makes those under him/her feel at ease when talking with him/her.	.86	.68	−.07	.17	.10	−.07
43. He/she is friendly and can be easily approached.	.82	.71	−.05	−.02	.05	−.05
44. He/she puts suggestions that are made by people under him/her into operation.	.87	.55	.14	.11	.26	.14
47. He/she gets the approval of his/her people on important matters before going ahead.	.62	.49	.05	−.02	.39	.05

[a] Sample A data from Fleishman (1953); Sample B data from Tscheulin (1973); Sample C data from present sample.

TABLE 2

COMPARISON OF SUPERVISORY BEHAVIOR DESCRIPTION QUESTIONNAIRE ITEM FACTOR LOADINGS ACROSS SAMPLE

Initiating structure items	Consideration (Factor I)			Structure (Factor II)		
	Sample[a]					
	A	B	C	A	B	C
2. He/she encourages overtime work.	.20	−.03	.09	.40	.43	.16
3. He/she tries out his/her new ideas on employees.	−.10	.35	.10	.42	.38	.33
6. He/she rules with an iron hand.	−.20	−.37	−.34	.58	.58	.51
8. He/she criticizes poor work.	−.18	−.18	−.15	.59	.63	.55
12. He/she talks about how much should be done.	−.20	−.09	−.32	.60	.55	.54
13. He/she encourages slow-working employees to greater effort.	−.17	.07	.33	.33	.74	.40
14. He/she waits for his/her employees to push new ideas before he/she does.	−.07	.07	−.02	.28	.33	.06
16. He/she assigns people under him/her to particular tasks.	.00	.08	.04	.26	.29	.37
20. He/she asks for sacrifices from his/her people for the good of the entire department.	.00	.09	.01	.46	.50	.41
23. He/she insists that his/her people follow standard ways of doing things in every detail.	.25	−.02	−.29	.72	.63	.49
27. He/she sees to it that people under him are working up to their limits.	−.17	.06	.16	.87	.73	.49
30. He/she offers new approaches to problems.	.36	.49	.63	.72	.30	.25
34. He/she insists that he/she be informed on decisions made by people under him/her.	.13	.11	−.03	.51	.57	.40
37. He/she lets others do their work the way they think best.	−.17	.44	.40	−.33	.20	−.32
38. He/she stresses being ahead of other work groups.	.03	−.16	−.10	.34	.42	.40
41. He/she "needles" people under him/her for greater effort.	.17	−.06	−.51	.50	.67	.48
42. He/she decides in detail what shall be done and how it shall be done.	.37	−.05	.12	.63	.49	.43
45. He/she emphasizes meeting of deadlines.	.10	.04	.17	.68	.49	.44
46. He/she asks people who are slow to work harder.	−.22	−.03	−.04	.40	.70	.57
48. He/she emphasizes the quantity of work.	.17	−.21	.01	.51	.69	.59

[a] Sample A data from Fleishman (1953); Sample B data from Tscheulin (1973); Sample C data from present sample.

(B) and the hospital sample (C) is exceedingly high, with coefficients of .95 and .93 for the Consideration and Structure factors, respectively. The low phi-coefficients between these factors, within and across samples, confirm the independence of these factors as separate dimensions of leader behavior.

CONCLUSION

The similarity of results of this present study (with a predominantly female population in a nonprofit environment) with those reported by Fleishman and Tscheulin lend increased confidence in the cross-sample stability of the SBD questionnaire and its use in the study of leadership in a variety of organizational contexts. In particular, the results of the three studies indicate that the construct validity of the SBD questionnaire is stable across time, sex differences, type of organization, culture, and analysis technique.

TABLE 3

FACTOR CONGRUENCY COEFFICIENTS FOR SUPERVISORY BEHAVIOR DESCRIPTION QUESTIONNAIRE

Sample and factor	Sample A		Sample C	
	Consideration	Structure	Consideration	Structure
Sample B				
Consideration	.91	.05	.95	−.05
Structure	.01	.81	−.14	.93
Sample C				
Consideration	.91	.02		
Structure	−.09	.80		

Note. Sample A data from Fleishman (1953); Sample B data from Tscheulin (1973); Sample C data from the present sample of Szilagyi-Sims.

REFERENCES

Fleishman, E. A. *Relationship between supervisory behavior and leadership climate.* Columbus, Ohio: Ohio State University Personnel Research Board, 1951.

Fleishman, E. A. The description of supervisory behavior. *Journal of Applied Psychology,* 1953, **37,** 1–6.

Fleishman, E. A. A leader behavior description for industry. In R. M. Stogdill & A. E. Coons (Eds.), *Leader behavior: Its description and measurement.* Columbus: Ohio State University, 1957.

Fleishman, E. A. *Manual for the Supervisory Behavior Description.* Silver Spring, Md.: American Institutes for Research, 1972.

Harman, H. H. *Modern factor analysis.* Chicago: University of Chicago Press, 1960.

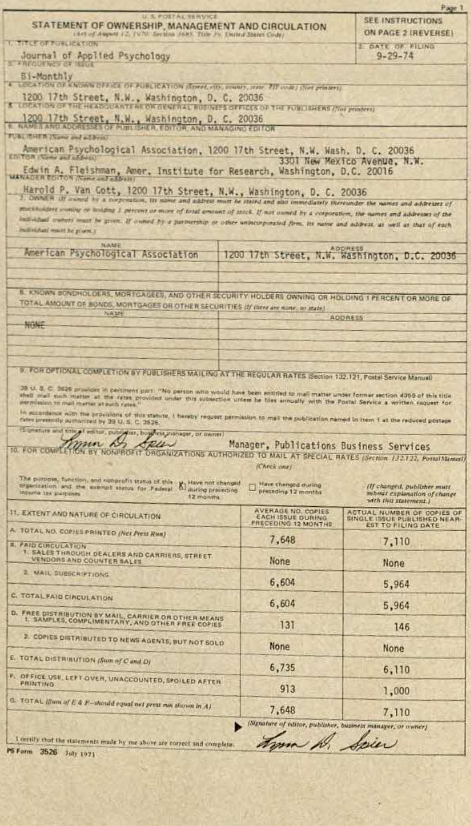
Nie, N., Bent, D., & Hull, C. *Statistical Package for the Social Sciences (SPSS).* New York: McGraw-Hill, 1970.

Stogdill, R. M., & Coons, A. E. (Eds.) *Leader behavior: Its description and measurement.* Columbus: Ohio State University, Bureau of Business Research, 1957.

Tscheulin, D. Leader behavior measurement in German industry. *Journal of Applied Psychology,* 1973, **57,** 28–31.

Wherry, R. J., & Winer, B. J. A method for factoring large numbers of items. *Psychometrika,* 1953, **18,** 161–179.

(Received September 13, 1973)



Page 1

U. S. POSTAL SERVICE
STATEMENT OF OWNERSHIP, MANAGEMENT AND CIRCULATION
(Act of August 12, 1970: Section 3685, Title 39, United States Code)

SEE INSTRUCTIONS ON PAGE 2 (REVERSE)

1. TITLE OF PUBLICATION: Journal of Applied Psychology
2. DATE OF FILING: 9-29-74
3. FREQUENCY OF ISSUE: Bi-Monthly
4. LOCATION OF KNOWN OFFICE OF PUBLICATION (Street, city, county, state, ZIP code) (Not printers): 1200 17th Street, N.W., Washington, D. C. 20036
5. LOCATION OF THE HEADQUARTERS OR GENERAL BUSINESS OFFICES OF THE PUBLISHERS (Not printers): 1200 17th Street, N.W., Washington, D. C. 20036
6. NAMES AND ADDRESSES OF PUBLISHER, EDITOR, AND MANAGING EDITOR
PUBLISHER (Name and address): American Psychological Association, 1200 17th Street, N.W. Wash. D. C. 20036
EDITOR (Name and address): Edwin A. Fleishman, Amer. Institute for Research, 3301 New Mexico Avenue, N.W. Washington, D.C. 20016
MANAGING EDITOR (Name and address): Harold P. Van Cott, 1200 17th Street, N.W., Washington, D. C. 20036

7. OWNER (If owned by a corporation, its name and address must be stated and also immediately thereunder the names and addresses of stockholders owning or holding 1 percent or more of total amount of stock. If not owned by a corporation, the names and addresses of the individual owners must be given. If owned by a partnership or other unincorporated firm, its name and address, as well as that of each individual must be given.)

NAME	ADDRESS
American Psychological Association	1200 17th Street, N.W. Washington, D.C. 20036

8. KNOWN BONDHOLDERS, MORTGAGEES, AND OTHER SECURITY HOLDERS OWNING OR HOLDING 1 PERCENT OR MORE OF TOTAL AMOUNT OF BONDS, MORTGAGES OR OTHER SECURITIES (If there are none, so state)

NAME	ADDRESS
NONE	

9. FOR OPTIONAL COMPLETION BY PUBLISHERS MAILING AT THE REGULAR RATES (Section 132.121, Postal Service Manual)

39 U. S. C. 3626 provides in pertinent part: "No person who would have been entitled to mail matter under former section 4359 of this title shall mail such matter at the rates provided under this subsection unless he files annually with the Postal Service a written request for permission to mail matter at such rates."

In accordance with the provisions of this statute, I hereby request permission to mail the publication named in Item 1 at the reduced postage rates presently authorized by 39 U. S. C. 3626.

(Signature and title of editor, publisher, business manager, or owner) Manager, Publications Business Services

10. FOR COMPLETION BY NONPROFIT ORGANIZATIONS AUTHORIZED TO MAIL AT SPECIAL RATES (Section 132.122, Postal Manual)
(Check one)

The purpose, function, and nonprofit status of this organization and the exempt status for Federal income tax purposes ☒ Have not changed during preceding 12 months ☐ Have changed during preceding 12 months (If changed, publisher must submit explanation of change with this statement.)

11. EXTENT AND NATURE OF CIRCULATION	AVERAGE NO. COPIES EACH ISSUE DURING PRECEDING 12 MONTHS	ACTUAL NUMBER OF COPIES OF SINGLE ISSUE PUBLISHED NEAREST TO FILING DATE
A. TOTAL NO. COPIES PRINTED (Net Press Run)	7,648	7,110
B. PAID CIRCULATION 1. SALES THROUGH DEALERS AND CARRIERS, STREET VENDORS AND COUNTER SALES	None	None
2. MAIL SUBSCRIPTIONS	6,604	5,964
C. TOTAL PAID CIRCULATION	6,604	5,964
D. FREE DISTRIBUTION BY MAIL, CARRIER OR OTHER MEANS 1. SAMPLES, COMPLIMENTARY, AND OTHER FREE COPIES	131	146
2. COPIES DISTRIBUTED TO NEWS AGENTS, BUT NOT SOLD	None	None
E. TOTAL DISTRIBUTION (Sum of C and D)	6,735	6,110
F. OFFICE USE, LEFT OVER, UNACCOUNTED, SPOILED AFTER PRINTING	913	1,000
G. TOTAL (Sum of E & F—should equal net press run shown in A)	7,648	7,110

I certify that the statements made by me above are correct and complete. ▶ (Signature of editor, publisher, business manager, or owner)

PS Form 3526 July 1971

Journal of Applied Psychology
1974, Vol. 59, No. 6, 771–772

RELATION BETWEEN FOREMEN'S LEADERSHIP ATTITUDES AND THE SKILL LEVEL OF THEIR WORK GROUPS

CHRIS CHING-YANG HSU[1] AND RICHARD R. NEWTON

General Motors Corporation, Detroit, Michigan

The relation between foremen's leadership attitudes, as measured by the Leadership Opinion Questionnaire, and the skill level of their work groups was studied. Subjects were 243 foremen and 66 general foremen in a manufacturing plant in the Midwest. The results indicated that the skill level of the foreman's work group is significantly related ($p < .01$) to his structure attitude but not to his consideration attitude toward these workers.

One situational variable which may be related to the leadership style of the supervisor is the skill level of the work group he supervises. Skilled workers have generally received more training and have gained higher prestige than unskilled workers. In addition, the jobs of skilled workers may call for more initiative and autonomy than those of unskilled workers. Supervisors of skilled workers may have less need to *structure* workers' roles and activities than supervisors of unskilled workers. Differences in skill level may also relate to the *consideration* shown by the supervisor, in terms of the rapport established, amount of two-way communication engaged in, etc.

The present study compares the leadership attitudes of structure and consideration, as measured by the Leadership Opinion Questionnaire (LOQ) developed by Fleishman (1969), for foremen supervising groups of skilled and unskilled workers.

METHOD

Subjects used in this study were 243 foremen, representing 85% of the foremen in the manufacturing area of a plant in the Midwest. Each foreman supervised about 25 hourly workers, averaged 42.2 years of age and 11.4 years of education.

Hourly workers were classified into skilled and unskilled groups. The former had served an apprenticeship leading to jobs as machinists, tool makers, electricians, etc. The latter had not served an apprenticeship and worked as assemblymen, painters, stockmen, etc. Foremen of skilled and unskilled worker groups did not differ significantly in mean age, education, or number of workers supervised.

The Leadership Opinion Questionnaire was administered at a location outside of the plant to general foremen and foremen in groups of approximately 80 persons as part of a manpower planning test battery. The subjects were told that the test results would be confidential and would not be available to their supervisors. The LOQ consists of 40 items, and it is purported to measure how a person thinks he should behave as a supervisor. The LOQ provides a score for structure and consideration (for detailed definitions of these leadership dimensions see Fleishman & Harris, 1962).

Since Fleishman (1953a) had previously shown that the leadership attitudes of the foreman's own boss might affect the attitudes of the foreman toward his workers, a check was made on the possible influence of this variable. The LOQ was administered to 66 general foremen who supervised the primary sample of foremen. Each general foreman supervised about four foremen and averaged 44.2 years of age and 12.3 years of education.

RESULTS

The average structure score of foremen was 51.08, and the consideration score was 50.78. The average structure score of general foremen was 52.58, and the consideration score was 54.30. General foremen have significantly ($p < .01$) higher consideration scores than foremen.

The correlations between structure and consideration scores for foremen and general foremen were $-.07$ and $-.10$, confirming the independence of these two scales reported earlier (Fleishman, 1953b).

Although general foremen of the unskilled group tend to have higher structure scores and lower consideration scores than general foremen of skilled workers, the differences were not statistically significant. Thus, both

---

[1] Requests for reprints should be sent to Chris Ching-Yang Hsu, 12-201 General Motors Building, Detroit, Michigan 48202.

TABLE 1

COMPARISON OF MEAN LEADERSHIP OPINION QUESTIONNAIRE SCORES FOR FOREMEN OF SKILLED AND UNSKILLED WORK GROUPS

LOQ scores	Groups	No. foremen	*M*	*SD*	*t* test
Structure	Skilled	41	47.88	7.56	2.92*
	Unskilled	202	51.72	7.69	
Consideration	Skilled	41	50.73	5.95	.02
	Unskilled	202	50.75	6.59	

Note. LOQ = Leadership Opinion Questionnaire.
* $p < .01$.

groups were reasonably well matched on this climate variable. However, foremen supervising the unskilled workers showed significantly ($p < .01$) higher mean structure scores than foremen of skilled workers (see Table 1). No difference in consideration score between foremen of the unskilled and skilled groups was found.

As a partial replication of the earlier Fleishman (1953a) findings, the attitudes of foremen working under general foremen with different leadership attitudes were compared. High consideration or structure attitudes were defined as those falling more than one half a standard deviation above the mean, while low consideration or structure attitudes were those falling more than one half a standard deviation below the mean.

Although the trends were in the direction of confirming the earlier findings (Fleishman, 1953a), that foremen's attitudes on consideration and structure are related to those of their own supervisors, the results of this supplementary analysis did not reach statistical significance.

CONCLUSION

The primary finding was that the skill level of the foreman's work group is a variable significantly related to the foreman's structure attitude toward his workers. On the other hand, the skill level of the foreman's work group was not related to the foreman's consideration attitude.

Korman (1966) suggested the development of "a systematic conceptualization of situational variance as it might relate to leadership behavior and a research program designed to test derivations from such a conceptualization . . . [p. 355]." The present results suggest one such situational variable which should be given further study in leadership research.

REFERENCES

Fleishman, E. A. Leadership climate, human relations training, and supervisory behavior. *Personnel Psychology*, 1953, **6**, 205–222. (a)

Fleishman, E. A. The description of supervisory behavior. *Journal of Applied Psychology*, 1953, **27**, 1–6. (b)

Fleishman, E. A. *Manual for Leadership Opinion Questionnaire*. (Rev. ed.) Chicago, Ill.: Science Research Associates, 1969.

Fleishman, E. A., & Harris, E. F. Patterns of leadership behavior related to employee grievances and turnover. *Personnel Psychology*, 1962, **15**, 43–56.

Korman, A. K. "Consideration," "initiating structure" and organizational criteria. *Personnel Psychology*, 1966, **19**, 349–361.

(Received November 19, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 6, 773–775

MECHANO-CONTENT ANALYSIS AS AN AID TO PREDICTING EFFECTS OF ADVERTISING MICROCOMMUNICATIONS

DONALD W. HENDON
Columbus College

A regression analysis of 204 posters was conducted to predict the percentage of respondents who remembered seeing an outdoor advertising message. The independent variables were 21 mechanical and 6 content elements; the dependent variables were recognition scores. The predictive validity of 20 equations was too low to be of value to advertising communicators. The correlations between predicted and actual scores were low (.03 to .41, $p < .05$) because the independent variables varied only slightly as a result of the outdoor medium's simplicity.

In June 1971, the Federal Trade Commission required all advertisers to substantiate every advertising claim. If the privilege of puffery has been removed from the advertiser's list of controllable variables, it is probable that content will be deemphasized, while graphics (the ads' mechanical variables) are emphasized. The graphics specialist who encodes the advertiser's original message in order to attract more attention, etc. will increase in relative importance to the advertiser. This specialist, who is usually the art layout designer, has been found to be out of touch with the receivers of advertising communications. Studies (Brinton, 1961; Tannenbaum, Jacobson, & Morris, 1964) have reported that the aesthetic judgments of graphic experts were not shared by ordinary readers.

The graphics specialist's job is to microcommunicate the advertiser's original macrocommunication to the receivers of his message by embellishing the original content with memorable or attention-getting mechanical features. The purpose of this study is to provide this member of the encoding team with guidelines to follow in the microcommunication process by indicating which mechanical variables aid in accomplishing the sender's goals—especially "recognition" of the message. Since both mechanical and content variables are used in the present context, a new term, "mechano–content analysis," is suggested as a substitute for "content analysis" in future studies where mixed variables are examined.

Outdoor poster ads were chosen because they are the simplest of all ad media, with the shortest and simplest verbal and visual messages. It was felt that if the simplest advertising communication channel were used, the results might be applied to more complicated ad media, possibly with a relatively high degree of predictive accuracy.

A review of the literature concerning the theory of the contribution of graphics to memory of ads revealed 53 separate studies suggesting 66 specific findings which would affect recall, recognition, or attention either positively or negatively (Hendon, 1973). For example, Twedt (1952) found in a study of trade journals that size of illustration, size and area of copy, number of colors, number of type sizes, number of illustration units, and number of copy units were significantly correlated with recognition scores in a positive direction.

This review led to the working hypothesis that in mass-communicated advertising messages via the outdoor medium, the Starch advertising research organization's recognition scores (Starch, 1962) are significantly correlated in a positive direction with the factors of brevity, simplicity, and boldness (or sensualness). Furthermore, a positive correlation is hypothesized between recognition and seven variables: (1) tallest product illustration letter height/poster height, (2) color area/poster area, (3) illustration area/poster area, (4) product illustration area/poster area, (5) product illustration area/illustration area, (6) product illustration area/illustration and product identification symbol (PID) areas, and (7) product illustration and PID areas/poster area. A negative correlation is hypothesized between recognition and 20 variables: (8) number of copy styles, (9) number of copy sizes, (10) average copy height/poster height, (11) tallest copy letter height/poster height, (12) number of copy lines, (13) number of copy units, (14) number of PID copy styles, (15) number of PID copy sizes, (16) average PID letter height/poster height, (17) number of colors, (18) number of illustration units, (19) PID area/poster area, (20) PID area/illustration and PID areas, (21) copy area/poster area, (22) number of copy letters, (23) number of copy words, (24) number of PID words, (25) number of product illustration words, (26) number of copy sentences, and (27) Robert P. Gunning's "Fog Index" (McKee, 1967). Variables 22–27 are the six content variables.

METHOD

Each of the independent variables were measured separately by two coders using color reproductions of the posters. A compensating polar planimeter was used to measure the areas which are included in the definitions of nine variables—2–7 and 19–21.

Each recognition score (dependent variable) was changed from the raw recognition score to a "copy performance index" (CPI) according to a technique developed by Starch (1962). Starch's scores were expressed in five different forms: (*a*) total CPI, (*b*) male CPI, (*c*) female CPI, (*d*) male CPI adjusted for inherent interest differences among products, and (*e*) female CPI similarly adjusted. All five CPI scores express the percentage of respondents who remembered seeing an outdoor ad.

There were 37 advertisers represented in three product categories: (*a*) gasoline/automobile accessories (89 posters), (*b*) food (68 posters), and (*c*) alcoholic beverages (47 posters). Thus, four equations were required for each of the five CPI forms—one for each product category and a fourth for all 204 posters grouped together.

The final sample, chosen on the basis of availability, covered 1956–1958 and 1961–1962. A cross-validation methodology was used by dividing each of the four product sets into four analysis subsets (Group A) and four validation subsets (Group B). A stepwise multiple regression and correlation technique was used to examine the relationship between recognition scores and the message elements in Group A to obtain predictive equations, and between Group B predicted and actual scores to test the predictive equations. The relationships between Group A predicted and actual scores and Group B scores and message elements were not measured.

A combination of factor and regression analyses was used to test the qualitative working hypothesis. Factor analysis was first used to summarize the information contained in a larger set of variables (27) in terms of another, smaller, less redundant set called "factors." Then, it was used in an indirect way to account for variations in the dependent variable in terms of the less redundant set of factors: If (*a*) all of the message elements which were significantly correlated with recognition scores in one direction were highly loaded on *only* one factor, and (*b*) no other message element was highly loaded on that same factor, then one could say that this particular factor was highly correlated with the Starch scores.

RESULTS

Because the aforementioned (*a*) and (*b*) did not occur, the working hypothesis was not supported. Five factors appeared in the all products and gas/auto categories: (*a*) copy content brevity and mechanical simplicity, (*b*) product illustration boldness, (*c*) visual boldness (*d*) copy area boldness, and (*e*) PID simplicity. Three more inclusive factors, brevity, simplicity, and boldness (or sensualness) appeared as modifying adjectives that described the five factors. Beyond these three more inclusive factors, the relationships between the message elements were unclear.

Using the first part of the cross-validation methodology (examination of the relationship between Group A independent variables and recognition scores), only the hypotheses corresponding to mechanical variables 3 and 7 were supported. These both highly loaded on the visual boldness factor; Twedt (1952) found this factor, which he labeled "pictorial-color." No other variables were so highly correlated with Group A recognition scores as were these (.59 and .51, respectively in the five gas/auto equations, $p < .01$). These were the only r values over .50.

Using the second part of the cross-validation methodology (examination of the relationship between Group B predicted and actual recognition scores), the correlations between predicted and actual Group B scores were quite low, ranging from .03 to .41. These r values were significantly greater than zero for only two sets of equations: all products and gas/auto ($p < .05$). Fairly large standard errors of the estimate occurred. The two strongest equations, both with r values of .41 ($p < .01$), were those for predicting the following:

(*a*) Percentage of recognition by men viewing gas/auto posters $= (1.15)\ [29.31 - .59$ (Variable 3) $- 3.01$ (Variable 8) $+ 1.08$ (Variable 11) $+ 1.39$ (Variable 15) $- .64$ (Variable 17) $- .89$ (Variable 21) $+ 1.22$ (Variable 25)]; (*b*) Percentage of recognition by women viewing gas/auto posters $= (.55)\ [67.70 - .96$ (Variable 4) $+ .60$ (Variable 7) $- 8.48$ (Variable 8) $- 7.64$ (Variable 24)].

These two equations were not applied to other ad media as originally intended because of low predictive validity found in the outdoor medium.

DISCUSSION

The following very general guideline is suggested for the graphics specialist: The larger the percentage of the poster's total area that is devoted to a visual illustration, especially that of the product and the PID (or signature), the higher the recognition scores of outdoor ads. However, since only 16% of the variance in Group B predicted and actual scores was accounted for by the variables in the two strongest equations, the creative interaction of the 27 variables, which cannot be measured, had to account for the remaining 84% of the variance, after one subtracted the influence of other variables which were not measured in this study. Therefore, it was not possible to quantify creativity, and the guide is probably of limited value.

Because of this limitation, the value of the present study lies elsewhere:

1. Very little of the variance in outcomes for a visual communications system for outdoor ads is accounted for by graphics decisions, because of the restriction of range in the numeric values of the independent variables. These numeric values were similar, since most poster ads are similar, having few words, simple layouts, etc. as a result of their unique viewing context. This finding has an implication for research methodology—Researchers should probably not begin their search for universals in the simplest environment possible and extrapolate and test their results in more complicated environments; the opposite approach is recommended. In the present context, the search for predictable outcomes for a communication system has to lie in a more complicated ad medium, one in which the messages transmitted by the medium would exhibit more differences than occurred in the outdoor sample.

2. McLuhan's (1970) thesis, "the medium *is* the message," was supported. There were intermedia differences that affected later recognition of ad messages carried by different media. The Twedt (1952) study and the present study had quite different equations predicting recognition scores. As Klapper (1960) suggested, a message is merely one influence, working amid other influences, in a total situation. Therefore, the pure message cannot be isolated and studied by itself. There will always be contamination from other forces, including the medium, which will affect the outcomes of a communication system.

REFERENCES

Brinton, J. E. The "feeling" of typefaces. *Communication Arts Magazine,* 1961, **3**, 43–45.

Hendon, D. W. How mechanical factors affect ad perception. *Journal of Advertising Research,* 1973, 13(4), 39–45.

Klapper, J. T. *The effects of mass communication.* New York: Free Press of Glencoe, 1960.

McKee, B. K. Readability formulas can aid good writing. *Public Relations Journal,* 1967, **23**(7), 8–11.

McLuhan, M., & Fiore, Q. *The medium is the massage.* New York: Bantam Books, 1970.

Starch, D. *Starch outdoor poster report #10-62. Interview period ending December 5, 1962: Automotive product group.* Mamaroneck, New York: Daniel Starch & Staff, 1962.

Tannenbaum, P. H., Jacobson, H. K., & Norris, E. L. An experimental investigation of typeface connotations. *Journalism Quarterly,* 1964, **41**, 65–73.

Twedt, D. W. A multiple factor analysis of advertising readership. *Journal of Applied Psychology,* 1952, **36**, 207–215.

(Received August 16, 1973)

Journal of Applied Psychology
1974, Vol. 59, No. 6, 776–778

EFFECTS OF ALTERNATIVE POSITIONING OF OPEN-ENDED QUESTIONS IN MULTIPLE-CHOICE QUESTIONNAIRES[1]

W. RUSSELL JOHNSON AND NICHOLAS A. SIEVEKING[2]
Vanderbilt University

EARL S. CLANTON III
Fisk University

Two studies investigated the effect of the placement of open-ended questions, relative to multiple-choice questions, on the responses elicited. Both studies found that positionings at both the beginning and the end of a predominantly multiple-choice questionnaire elicited significantly more responses than did either placement alone. When the two positions were used singly, the one at the beginning was found to elicit more discrete ideas and response categories. This was true for black samples and white samples and for different questionnaire topics and lengths.

Most questions asked in opinion research are multiple-choice because such items often are easier to answer, require less of the respondents' time, and permit direct quantification of answers. However, these questions are often supplemented with open-ended questions in order to elicit unanticipated responses.

Open-ended questions are frequently positioned among other questions wherever convenient or logical, without considering that responses to these questions could be influenced by their placement. Placement at the end of a questionnaire might elicit both the respondent's initial ideas and those suggested to him during the questionnaire. Alternatively, such placement might cause response attenuation due to fatigue. Although authors have given attention to item construction (e.g., Wood, 1960) and to the use and arrangement of questions in an interview (e.g., Kahn & Cannell, 1957), no studies of the effects of various placements of open-ended questions have been conducted. Due to the cost of administering and quantifying open-ended questions, it seems advisable to determine which placements produce more responses. The present investigation compared responses to open-ended questions placed at the beginning of written questionnaires and at the end, after intervening multiple-choice items. The problem of open-ended question placement within an interview was not addressed.

METHOD

To investigate the generality of positioning effects, two separate studies were conducted. These involved different questionnaire content, different experimenters, and different subject samples. The media study, so named because its items related to the news media, was given in class by a black experimenter to 110 black high school and college students. An open-ended question was placed at *both* the beginning (Item 1) and the end (Item 62) of the questionnaire ("first and last" condition), after 60 multiple-choice items dealing with the quality of local news media. The question asked: "What changes would you like to see in the news presentations by television, radio, and newspapers?"

The second study, the counseling study, included items dealing with various attitudes about counseling and counseling methods. The counseling study questionnaire was administered by a white experimenter during two introductory psychology classes to 156 white college students. The open-ended question asked: "What questions about yourself would you like to have answered if counseling were made available to you?" There were two conditions in the counseling study. In one of these, 75 randomly chosen subjects received the open-ended question at *both* the beginning (Item 1) and the end (Item 18) of the questionnaire (first and last condition). The intervening 16 multiple-choice items dealt with less personal matters such as attitudes toward various methods of counseling. In the second condition, 81 subjects received the open-ended question in the last position (Item 18) only ("last only" condition).

---

[1] The authors wish to express their thanks to Elaine Johnson, who served as one of the judges, and to Lynn Cocke, who assisted with data analyses and preparation of the manuscript.

[2] Requests for reprints should be sent to Nicholas Sieveking, Interuniversity Psychological and Counseling Center of Vanderbilt University, George Peabody College and Scarritt College, 300 Oxford House, 1313 21st Avenue South, Nashville, Tennessee 37212.

TABLE 1

MEAN NUMBER OF DISTINCT IDEAS GIVEN PER RESPONDENT

Study	Open-ended question position											
	First (all subjects)			Last (all subjects)			Total: First and Last (all subjects)			Last (Subjects giving additions)		
	n	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Media												
First and last	110	2.33	1.48	110	1.31	3.25	110	3.64	2.27	57	2.53	1.60
Counseling												
First and last	75	1.80	1.10	75	.51	.74	75	2.31	1.40	29	1.26	.60
Counseling												
Last only	—	—	—	—	—	—	81	1.36	.96	—	—	—

Subjects who received the open-ended question at the beginning (Item 1) were asked not to alter their responses to that question once they had begun answering the multiple-choice items. Since those subjects were not aware that the open question asked at the beginning would be repeated, the first positioning within the first and last condition also constituted a "first only" condition.

Two judges, blind to question positions, evaluated each subject's responses to the open-ended questions in terms of the number of separate thoughts expressed, producing a *distinct ideas* score for each respondent. The judges also sorted responses into broader *response categories* which were abstracted from subjects' answers by grouping related distinct ideas. Independent interjudge agreement for distinct ideas was 90% in the media study and 94% in the counseling study. Corresponding percentages for response categories were 69% and 62%. Disagreements were settled by discussion between judges.

RESULTS AND DISCUSSION

Distinct ideas. Subjects in both studies added significantly ($t = 5.06$, $df = 218$, $p < .001$, media study; $t = 2.49$, $df = 148$, $p < .01$, counseling study) to the number of distinct ideas given to the open-ended question in the first position when they were given a second opportunity to answer the same question (see Table 1). The additional distinct ideas given by media study subjects averaged 1.31 and those given by counseling study subjects averaged .51. These averages are based on each entire subject sample, though not all subjects offered additional responses when given a second opportunity. Of the media study subjects, 52% did give additional responses, and these subjects averaged 2.53 additional distinct ideas per response. The corresponding average for the 39% of counseling study subjects who gave additional responses was 1.26.

In answering the question in the last only position, counseling study subjects gave significantly ($t = 2.68$, $df = 154$, $p < .01$) fewer distinct ideas per subject than when asked the same question in the first position and significantly ($t = 4.95$, $df = 154$, $p < .001$) fewer than when asked the same question twice. Thus, in terms of eliciting a maximum number of distinct ideas per subject, the first only position appeared more productive than the last only position, and the first and last position was more productive than both first only and last only positions.

Response categories. The same pattern of results emerged with response categories as did with comparisons of distinct ideas. The 400 and 175 distinct ideas given by subjects in the media and counseling studies, respectively, were placed into 43 and 18 categories. Both studies demonstrated that respondents used significantly more categories per subject in the first and last positioning than in either the first only ($t = 6.46$, $df = 218$, $p < .001$, media study; $t = 2.04$, $df = 148$, $p < .05$, counseling study) or the last only position ($t = 3.90$, $df = 154$, $p < .001$) (see Table 2). Also, a first only position elicited significantly ($t = 2.20$, $df = 154$, $p < .05$) more response categories than a last only position.

It might have been expected that repetition of an open-ended question would elicit repetition of previous ideas or at least of ideas similar enough to be in the same category; however, this was not the case. Few respondents used categories in

TABLE 2

MEAN NUMBER OF RESPONSE CATEGORIES USED PER RESPONDENT

Study	Open-ended question position											
	First (all subjects)			Last (all subjects)			Total: First and Last (all subjects)			Last (Subjects giving additions)		
	n	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Media												
First and last	110	1.81	1.03	110	1.06	1.25	110	2.87	1.36	61	1.93	1.10
Counseling												
First and last	75	2.23	1.26	75	.51	.85	75	2.73	1.71	24	1.58	.82
Counseling												
Last only	—	—	—	—	—	—	81	1.79	1.23	—	—	—

answering a repeated question that they had used in answering it the first time. Of the subjects in the first and last condition, 47% and 21% repeated averages of 1.61 and 1.13 categories per subject in the media and counseling studies, respectively. The remaining subjects repeated no categories. Of the media study subjects, 55% used additional categories on the repeated question; the average number of new categories for these subjects was 1.93. Similarly, 32% of counseling study subjects used an average of 1.58 new categories. Very few of these subjects also repeated categories.

These studies offer evidence that if an open-ended question is included only once, the best placement for it is at the beginning of the questionnaire. However, repetition of an open-ended question elicited more distinct ideas per subject and more response categories per subject than did either single placement. The generality of these results was suggested by their occurring with black subjects and white subjects and with different questionnaire topics and lengths. Since the costs of developing response codes and categories should be similar whether a question is given once or repeated, the return per investment should be greater if open-ended questions are asked twice.

REFERENCES

Kahn, R. L., & Cannell, C. F. *The dynamics of interviewing*. New York: Wiley, 1957.

Wood, D. A. *Test construction: Development and interpretation of achievement tests.* Columbus, Ohio: Charles E. Merrill, 1960.

(Received October 16, 1973)

RESEARCH